University of Alberta

Counterfactuals

by

Vladan Djordjevic ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the

requirements for the degree of Doctor of Philosophy

Department of *Philosophy*

Edmonton, Alberta

Fall 2005

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-08631-9

Your file Votre référence
ISBN:
Our file Notre reterence
ISBN:

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

In this thesis I try defend a version of a pragmatic theory of counterfactuals. The main points and results of the thesis are:

- Usual interpretations of Goodman have consequences that neither Goodman nor most of the interpreters would accept
- Showing the relation between the metalinguistic theory and various possible worlds semantics
- A counterexample to Lewis's notion of cotenability
- Showing the impossibility of theories of Lewis's type – total ordering possible worlds semantics based on absolute similarity or selection function
- Rejecting antecedent-relative and antecedent-and-consequent relative semantics
- A critique of Warmbrōd's context-relative semantics
- Comments on von Fintel's dynamic semantics
- Showing some features of context that influence the truth values of counterfactuals
- Pointing to the problem of the fragility of propositions, showing its relevance to counterfactuals, and proposing a solution.

- Showing that counterfactuals are context-dependent and context-influencing in such a way that logic of counterfactuals, in the usual sense of 'logic', cannot be made; a pragmatic theory is needed.
- Proposing a maximal change context-relative theory of counterfactuals; this is a version of a pragmatic theory; it borrows from Gabbay, Warmbrōd and von Fintel.

Table of Contents

1. Preface p. 1
2. Goodman p. 11
3. Stalnaker p. 21
4. Lewis p. 37
5. Absolute similarity p. 56
6. Relative similarity p. 70
    - 6.1 Antecedent-relativity p. 70
    - 6.2 Gabbay's antecedent-and-consequent relativity p.77
    - 6.3 CSO p. 79
    - 6.4 Warmbrõd's context-relativity p. 92
    - 6.5 Relative theories and the metalinguistic theory p. 97
7. A context-relative theory of counterfactuals p. 103
    - 7.1 A modification of Warmbrõd's theory p. 106
    - 7.2 von [illegible]intel's dynamic semantics p.115
    - 7.3 On four aspects of context-dependency p. 123
    - 7.4 A modification of von Fintel's theory p. 141
    - 7.5 A pragmatic theory of counterfactuals p. 145

Bibliography p. 155

List of Figures

Figure 1 p. 39

Figure 2 p. 45

Figure 3 p. 50

1. Preface

Almost anything interesting anybody said about counterfactuals is disputed. My intention in this thesis is not to address all these issues, but only those about which I had something original, and, I hope, interesting to say. My opinion about issues that I will *not* discuss is in accordance with well-established, well-defended views held by lots of philosophers (although they are rejected by lots of other philosophers as well). The purpose of this introduction is to list these views that will be *assumed* throughout this thesis. After that I will tell what the thesis is about, in the form of a longer abstract.

My assumptions are:

1.1 Counterfactuals have truth values.

1.2 The terms 'counterfactuals' and 'indicative' are not good names for the two kinds of conditionals they are supposed to refer to. I use them only because they became common long ago, and no new names have become widely accepted in the meantime. My assumption here is that this will not lead to confusion.

1.3 'Even-if' conditionals deny that there is a connection between the antecedent and the consequent, which is why they are neither indicative conditionals nor counterfactuals (both of which claim the existence of some kind of a connection between the antecedent and the consequent).

1.4 No backtracking. A counterfactual conditional 'Had A been the case then C would have been the case' is backtracking if C comes temporarily before A; C is a cause or a reason that would have to have obtained for A to be the case. I find backtracking counterfactuals either confusing or plainly false. The corresponding indicative conditionals look to me very clear. What is intended to be said using backtracking counterfactuals can be said with the corresponding indicative conditionals, without any confusion.

1.5 For any rule of inference involving counterfactuals there is some sane person who thinks that the rule is invalid. I am glad to be able to offer an explanation of that fact and still explain why we can talk about good and bad reasoning involving counterfactuals.

These were the main assumptions, and my work begins in Chapter 2 on Goodman. I apologize for the boring start – it will be a picky analysis of what Goodman said and what he did not say. However, it will soon become clear that the basic idea of the metalinguistic theory is essential for this thesis, and thus requires my special attention. Each chapter will have an important part that will relate the metalinguistic theory to some possible worlds semantics. The main points and results of the thesis are:

- Usual interpretations of Goodman have consequences that neither Goodman nor most of the interpreters would accept
- Showing the relation between the metalinguistic theory and various possible worlds semantics
- A counterexample to Lewis's notion of cotenability
- Showing the impossibility of theories of Lewis's type – total ordering possible worlds semantics based on absolute similarity or selection function
- Rejecting antecedent-relative and antecedent-and-consequent relative semantics

- A critique of Warmbrōd's context-relative semantics
- Comments on von Fintel's dynamic semantics
- Showing some features of context that influence the truth values of counterfactuals
- Pointing to the problem of the fragility of propositions, showing its relevance to counterfactuals, and proposing a solution.
- Showing that counterfactuals are context-dependent and context-influencing in such a way that logic of counterfactuals, in the usual sense of 'logic', cannot be made; a pragmatic theory is needed.
- Proposing a maximal change context-relative theory of counterfactuals; this is a version of a pragmatic theory; it borrows from Gabbay, Warmbrōd and von Fintel.

Here are detailed contents:

Counterfactuals

(Contents)

1. Preface p. 2

2. Goodman p. 12

Goodman's metalinguistic theory of counterfactuals is often not interpreted carefully. Often interpretations of Goodman give counterfactuals some formal properties that neither Goodman nor the interpreter would accept. I try to present a correct interpretation, which will later be compared to all other theories mentioned in this thesis. What we call today the metalinguistic theory says that a counterfactual A→C is true iff the argument {A, $B_1, \dots, B_n\} \vdash C$ is valid. The B's are the so-called *background* propositions. They are some truths cotenable with A. We should distinguish the B's that are *relevant* for a given conditional, and irrelevant ones. The *usual interpretation* (UI) of the metalinguistic theory says that both relevant and irrelevant background propositions are included among the B's. According to Lewis *some*, and according to Nute *all*, irrelevant cotenable truths are to be included among the B's. There is no basis in Goodman's paper to claim which one, if any, of these interpretations is correct.

3. Stalnaker p. 22

Stalnaker's semantics is presented and its philosophical ambitions are compared to the ambitions of the metalinguistic theory. Conditional excluded middle (CEM), the distinctive feature of Stalnaker's semantics for conditionals, is commented on. It is then shown that CEM is validated also by Nute's version of UI, although that was not intended to be the case. Still Nute's UI is not equivalent to Stalnaker's theory, because it does not validate CSO, one of Stalnaker's axioms. It has been generally accepted that in interpreting Goodman's theory it was not important whether the irrelevant B's are to be included, and if so, which ones. In this chapter I show that this was wrong, since the choice of the irrelevant B's influences the formal properties of counterfactuals.

4. Lewis p. 38

David Lewis explained what he thought to be the relation between his and the metalinguistic theory, and defined cotenability in terms of his similarity relation. I give a counterexample to his notion of cotenability. Then I show that the two theories are not related in the way Lewis thought they were, and I argue that this is a problem for Lewis's semantics, not for the metalinguistic theory.

5. Absolute similarity p. 57

Possible worlds semantics use as a primitive notion a similarity relation or a selection function (or can be reformulated that way). According to the properties of that notion Nute classified possible worlds semantics as *minimal*, *small* or *maximal change* theories. I add another classification that distinguishes theories based on an *absolute* or on a *relative* similarity or selection function. The purpose of that primitive notion is to somehow separate the worlds that are important from those that are unimportant when we evaluate the truth value of a counterfactual. The meaning of 'important' worlds depends on the theory (the closest antecedent worlds, antecedent worlds that are similar enough, etc. etc.). In any case, I argue that a part of the meaning of 'important' *must* allow the following interpretation: the relevant background propositions are true at all the important worlds. Then I argue that Lewis's theory does not satisfy this condition. The same holds for any theory of that kind – minimal or small change total ordering semantics based on an absolute similarity relation or selection function.

6. Relative similarity p. 71

6.1 Antecedent-relativity p. 71

If we cannot use the same similarity relation or selection function for all counterfactuals, as the arguments from the previous chapter imply, this suggests that we should use a relative notion of similarity. The most popular such notion is the *antecedent*-relative similarity. I examine a version of that notion and I conclude that it is not 'relative enough'. This suggests the need for another kind of relativity.

6.2 Gabbay's antecedent-and-consequent relativity p.78

Gabbay proposed semantics based on an *antecedent-and-consequent*-relative selection function. Despite many intuitively attractive features of that notion, the final result is disastrous – the semantics based on it makes invalid almost all the rules of inference that we use in natural language.

6.3 CSO p. 80

On the formal level, the dispute about the absoluteness and relativity of similarity has, among its main tasks, to justify the acceptance or rejection of CSO. That formula: $(A\to B)\wedge(B\to A)\supset(A\to C)\equiv(B\to C)$, which is an axiom of every absolute theory, is invalid in most relative theories. To decide who is right in that regard, I examine various counterexamples to CSO. It turns out that both absolute and relative theories have good reasons for their views, but neither can explain all the problems involved in the discussion on CSO. Also, the discussion has the purpose of introducing new questions about counterfactuals that were rarely or never discussed in the literature on conditionals. The questions involve backtracking, a certain kind of context-dependency (namely, dependency of what was previously said in a conversation), and the fragility of propositions. An attempt to solve these questions is left for the next chapter.

6.4 Warmbrōd's context-relativity p. 93

Here I present Warmbrōd's theory and some motives for introducing this kind of relativity.

6.5 Relative theories and the metalinguistic theory p. 98

Chapter 6 finishes with a comparison of the metalinguistic theory and relative semantics. Nute's UI turns out to be a metalinguistic counterpart of the version of the antecedent-relative possible worlds semantics that I discussed in 6.1. Lewis's UI corresponds to a different version of the antecedent-relative theory, and a counterpart for Gabbay's semantics would be a version of the metalinguistic theory where the set of B's is restricted *only* to the relevant background propositions.

7. A context-relative theory of counterfactuals p. 104

7.1 A modification of Warmbrōd's theory p. 107

Warmbrōd proposed a minimal change semantics based on *context*-relative similarity. I argue that the involvement of context-relativity is a big improvement, but that the notion of similarity leads his theory to some problems. To avoid the problems I suggest the following modification. Instead of the notion of similarity, which I find misleading, I use a selection function similar to Gabbay's. Next, instead of Warmbrōd's 'normality condition', which tells us when counterfactuals belong to the same context, I propose that counterfactuals 'go together' (i.e. can be evaluated in the same model) according to the following restriction. For any set of counterfactuals, if all the background propositions for all counterfactuals from the set are cotenable with each antecedent appearing in the set, than the counterfactuals from the set go together. A rule of inference applies only if all counterfactuals appearing in the rule go together. Although I think that the modified Warmbrōd's theory is good for many cases, it cannot explain all cases, which is why we need to make more modifications with the aid of von Fintel's theory.

7.2 von Fintel's dynamic semantics p.116

Here I present von Fintel's theory of counterfactuals. The theory assumes a major change of the standard views on meaning and the semantics-pragmatics border. Truth conditions are much more context-dependent than it was usual in formal semantics – namely the truth value of a counterfactual depends on what has previously been said in the conversation. Technically this kind of context-dependency is nicely captured in dynamic logic.

7.3 **On four aspects of context-dependency** **p. 124**

In this section I discuss four aspects of context that (can) have influence on the truth values of counterfactuals: 1) what has previously been said in conversation, 2) the state of affairs in the spatio-temporal relevant region 3) fragility of propositions, 4) principle of charity. Put more technically, the topic is how the selection function is influenced by the four aspects of context.

7.4 **A modification of von Fintel's theory** **p. 142**

Von Fintel's theory is based on some elements of minimal change absolute theories. I replace those with the elements of Gabbay's theory.

7.5 **A pragmatic theory of counterfactuals** **p. 146**

Based on the arguments from section 7.3, I argue that counterfactuals are context-dependent and context-influencing in such a way that no conditional logic stronger than Gabbay's 'entailment free' logic is possible. At the same time, the fact that in ordinary language we do use lots of rules that are invalid in Gabbay's logic suggests that we often assume some restrictions on the potential that counterfactuals have to influence the context. I describe two such restrictions: one captured by my modification of Warmbrod's theory, the other by my modification of von Fintel's theory.

Bibliography p. 156

2. Goodman

Since the metalinguistic theory of counterfactuals plays a big role in this thesis, I will start with reminding the reader of the classical formulation of the theory made by Goodman. I believe that Goodman's theory is often not interpreted carefully, and I will attempt to give a correct interpretation. Having realized that he was not able to finish his project, Goodman abandoned his theory and left it in quite a mess. That left room for different interpretations. However, it is not rare that interpretations of Goodman give counterfactuals some formal properties that neither Goodman nor the interpreter would accept. I try to show that in this and the next chapter.

The term 'metalinguistic' was first used by David Lewis[1]. According to the metalinguistic theory, he said, a counterfactual conditional A→C is true (Goodman[2]) or assertable (Rescher[3]), iff A, in conjunction with a set of further premises, implies C. In other words, A→C holds iff there is a valid argument

$$\frac{A, B_1,\ldots,B_n}{C} \qquad (2.1)$$

where $B_1,\ldots,B_n$ are true contingent propositions. Either the counterfactual is a sentence meaning that there is such an argument, or (Mackie[4]) that the conditional is an elliptical presentation of such an argument, i .e. that a counterfactual sentence is neither true nor false, but rather valid or invalid. Hence the name 'metalinguistic theory'[5]. Throughout this paper I deal only with the metalinguistic theory that provides truth conditions for counterfactuals and only with Goodman's version of it.

---

[1] Lewis 1973 p.65.

[2] Goodman 1947

[3] Rescher 1964

[4] Mackie 1962

[5] Under the authority of Lewis this name become common. Bennett thinks that we still don't have a good label for this kind of theories. Cf Bennett 2003. p.303. He uses the term 'support theory'. Hansson (1995) uses 'derivability theory'.

The main task for this type of theory is to determine what is to be included among the relevant auxiliary or background propositions $B_1,...,B_n$. Judged by the modern standards, Goodman's attempt to solve this problem was very ambitious. He wanted to define the set of B's in precise and well-defined terms. At one point in his paper Goodman suggested the following tentative definition:

> (TD) $A \rightarrow C$ is true iff: (α) there is some set S of true sentences such that S is compatible with C and with $\neg C$, and such that $\{A\} \cup S$ is self-compatible and leads by law [of nature + logic] to C; while (β) there is no set S' such that S' is compatible with C and with $\neg C$, and such that $\{A\} \cup S'$ is self-compatible and leads by law to $\neg C$.[6]

Let A=the match m is struck, C= the match m lights, and let both A and C be false. Let the elements of the set S of true relevant propositions be: B_1=m is dry, B_2=m is well made, B_3=oxygen enough is present, etc. This is the example Goodman used to test TD. To make things simpler we can suppose that S contains the relevant laws of nature, e.g. B_4 = All dry, well made matches light when struck in the presence of oxygen. Then S $=\{B_1, B_2, B_3, B_4\}$. That way the corresponding argument 2.1 is valid only in virtue of laws of logic. We want our definition to consider true the proposition

---

[6] "The Problem of Counterfactual Conditionals" in F. Jackson (ed.) 1991. pp.16-7. Originally published in Goodman 1947.
The labels α and β do not appear in the original. I put them following Parry 1957. Note that Goodman uses 'S' both for the set of B's and for the conjunction of B's. Cf. p. 14.

2.2 Had m been struck, it would have lit. ($A \rightarrow C$)

Indeed, it seems to satisfy TD. However, the problem is that $A \rightarrow \neg B_i$, for any i, also seems to satisfy TD. For example

2.3 Had m been struck, it would have been wet. ($A \rightarrow \neg B_1$)

The set of true relevant propositions for this conditional is $\{\neg C, B_2, B_3, B_4\}$ (call it S_1). S_1 and A together entail $\neg B_1$. Thus TD is wrong, because intuitively it is obvious that 2.2 is true and 2.3 false. We need a definition that would make the right choice between 2.2 and 2.3.

It is worth noting that the choice between 2.2 and 2.3 is related to another essential feature of counterfactuals, namely that they should be distinguished from indicative conditionals. 2.2 is true and all of $A \rightarrow \neg B_i$ are false for any i, while in their indicative mood it should be the opposite – 2.2 should be false and at least one of $A \rightarrow \neg B_i$ should be true. For example:

(2.2i) If m was struck, it lighted.

(2.3i) If m was struck, then it was wet (or not well made, or there was not enough oxygen).

We see a new match m that never lighted, so it did not light even if it was struck. Thus 2.2i is false. And since it never lighted, then if it was struck, it must be that it didn't light either because it was wet or not well made or... etc., as 2.3i says.

Actually, Goodman was wrong that TD makes 2.2 and 2.3 true. In fact, TD makes them both false. Take $\{A \supset C\}$ as S and $\{A \supset \neg C\}$ as S' and the condition β would not be satisfied ('$\supset$' is the material implication). This problem was discovered by Parry[7]. Goodman admitted that Parry was right[8], and in later reprints of his 1947 essay we can see the footnote:

> Since this essay was first published, W. T. Parry has pointed out that no counterfactual satisfies this formula [TD]; for one could always take $\neg(A\wedge\neg C)$ as S, and take $\neg(A\wedge C)$ as S'. Thus we must add the requirement that neither S nor S' follows by law from $\neg A$.[9]

(Note that $\neg(A\wedge\neg C)$ is equivalent to $A\supset C$ and $\neg(A\wedge C)$ to $A\supset\neg C$. Goodman apparently has in mind only 'real' counterfactauls with a false antecedent.)

However, the main problem with TD, which cannot be removed by the cited additional requirement, is that $\neg C$ *would not have been* true had A been true, which means that the set S_1 would not have been a set of true propositions had A been true. In other words, $\neg C$ is not *cotenable* with A. It is implicit (but still very clear) in Goodman's paper that he would define cotenability thus: B is cotenable with A iff $\neg(A\rightarrow\neg B)$. To improve TD, Goodman suggested that $A\wedge S$ should not only be self-compatible, but S should be cotenable with A as well. As I mentioned in footnote 6 on page 13 above, Goodman used 'S' to denote both the set of B's and the conjunction of all the B's. Thus S being cotenable with A means either that the conjunction of the B's is cotenable with A, or that each member of S is cotenable with A. Goodman never felt the need to be more precise on this point.

---

[7] Cf. Parry 1957.

[8] Cf. Goodman 1957.

[9] This is footnote 6. I don't know when it was first added to the 1947 essay. It appears in Goodman 1984 and later reprints.

What are the final truth conditions Goodman proposed? We cannot tell. Instead of putting them explicitly, Goodman said:

> Returning now to the proposed rule, I shall neither offer further corrections of detail nor discuss whether the requirement that S be cotenable with A makes superfluous some other provisions of the criterion; for such matters become rather unimportant besides the really serious difficulty that now confronts us.[10]

The difficulty is that that he could not define counterfactuals without the notion of cotenability, while cotenability is defined in terms of counterfactuals. Being unable to avoid this circularity, Goodman thought that his whole project was a failure, and wouldn't bother any more with technical details.

The facts that Goodman did not offer a final formulation of the truth conditions, that TD is quite long, that it would be even longer with the addition of Parry's improvements and the notion of cotenability, and that those additions would make parts of TD redundant, left room for different interpretations. Very often Goodman's theory is presented with most parts of his definition ignored, and the rest slightly changed. Authors do that probably thinking that the ignored parts are redundant or even wrong, and that the changes they made either improve Goodman's definition, or express explicitly what they take was only implicit in Goodman's paper. Some of those interpretations are simply too inaccurate to be ascribed to Goodman. For example:

---

[10] Goodman 1991 p. 19.

> Intuitively, S is to consist of sentences which (i) are true and (ii) would also have been true, if contrary to fact, ϕ [the antecedent] had been true. The second condition, ... , Goodman referred to as the cotenability of S with ϕ.[11]

Today the usual interpretation of Goodman says that

(UI) A→C is true iff A, together with a set S of true premises, each of them cotenable with A, entail C.

UI again has different versions. At one point Lewis says that the set of cotenable premises is *finite*[12]. According to Nute's interpretation[13] S contains every truth cotenable with A, which means that S is *infinite*. In that case the cardinality of S is at least $\aleph_1$, or as big as our formal language permits it to be. To distinguish these two interpretations, one saying that S is finite, the other that S is infinite, let us call them Lewis's UI and Nute's UI. (Please understand these as labels that I find convenient to use at this point. I don't mean to imply that Lewis's UI is the exact way he reads Goodman – Lewis did not give an explicit and complete interpretation. Nute's UI, on the other hand, seems to be what Nute ascribes to Goodman.)

Let us note first that from Goodman's paper we cannot tell whether he thought that S was finite or not. On one hand, he does speak of a conjunction of the background propositions.[14] This goes in favor of Lewis's version, since conjunctions, like other classical logic formulae, are finite by definition. On the other hand, at the beginning of his section on relevant background propositions Goodman said:

---

[11] Turner 1981. p. 453.

[12] Cf. Lewis 1973. p. 57

[13] Cf. Nute and Cross p.5.

[14] Goodman 1991 p. 14, and on several other places throughout the text.

> It might seem natural to propose that the consequent follows by law from the antecedent and a description of the actual state-of-affairs of the world, that we need hardly define relevant conditions because it will do no harm to include irrelevant ones. But...[15]

Then Goodman explains why this won't work (because the negation of the antecedent is part of the description of the actual state-of-affairs). After that Goodman never mentioned whether S should contain irrelevant conditions or not. If S contained a description of the world, it would be infinite. Goodman rejected this suggestion, but he did not discuss whether his TD requires that S contains *only* the relevant background proposition (e.g. B_1 – B_4 in the example with the match m), or it allows some irrelevant propositions to enter S (and if so, how many). Thus we cannot say that either of the two versions of UI is a wrong interpretation of Goodman in regard of the cardinality of S, since he was not clear on that point. Let us express the differences between the two versions this way: besides the relevant propositions, the set of background propositions contains *some* (Lewis's UI), or *all* (Nute's UI) irrelevant truths cotenable with the antecedent.

UI is much shorter than TD. What happened to all those conditions Goodman talked about? Let us consider them one by one and see if they are justifiably omitted or changed.

---

[15] Ibid. p. 13.

UI ignores Goodman's tendency to separate the laws of nature from (other) contingent truths from S. I will do the same throughout this thesis, because it makes things simpler and no harm comes out of it.

In UI there is only one criterion for a true proposition to be included in S – being cotenable with A. Obviously Goodman thought that not all cotenable truths are to be in S, unless they satisfy the rest of the conditions. Thus he would distinguish the notions of 'truth cotenable with the antecedent' and 'background proposition'. The former includes the latter, but not the other way around. Unless we prove that all other conditions are redundant (or wrong), we should keep the distinction between the two notions.

Let us see first what happened to the negative condition β. Goodman's immediate reason to add this condition to α was to avoid having true both 2.7 and 2.8

2.4 Jones is not in South Carolina

2.5 Jones is not in North Carolina

2.6 North Carolina plus South Carolina is identical with Carolina

2.7 If Jones were in Carolina, he would be in South Carolina

2.8 If Jones were in Carolina, he would be in North Carolina [16]

Goodman said that "Clearly it will not help to require only that for *some* set S..."[17] the condition α holds. Because sets {2.4, 2.6} and {2.5, 2.6} can serve as the set S for 2.8 and 2.7 respectively, so both 2.7 and 2.8 satisfy α. Neither satisfies β, so both are false according to TD, which is intuitively the right result. That was Goodman's reason to include β in TD. However, after Goodman pointed out that TD had to be improved with the notion of cotenability, we can see that the 2.7-8 example is solved as well: 2.7 and 2.8

---

[16] Ibid p. 15.

[17] Ibid

cannot both be true. If 2.7 is true, then 2.5 is not cotenable with the antecedent 'If Jones were in Carolina'; therefore 2.5 cannot be included in S, and hence 2.8 is false. If 2.8 is true, then 2.4 is not cotenable with the antecedent, so now 2.7 is false. Thus Goodman's paper does not give us any reason to keep β.

The additional requirement that Goodman added under Parry's influence as footnote 6, however, cannot be ignored. After Goodman said that cotenability has to be involved in TD, there is no danger any more that no counterfactual would come up true. If $A \to C$ is true and A false, both $A \supset C$ and $A \supset \neg C$ are compatible with A, but $A \supset \neg C$ is not cotenable with A. However, another problem immediately arises. If both $A \to C$ and $A \to \neg C$ are (intuitively) false, then both $A \supset C$ and $A \supset \neg C$ are cotenable with A. According to UI both $A \supset C$ and $A \supset \neg C$ are in S. Then $S \cup \{A\}$ is inconsistent, which makes all counterfactuals with the antecedent A true, contrary to our assumption that $A \to C$ and $A \to \neg C$ were false. Therefore UI is not a correct interpretation of Goodman, and we should keep the distinction between the notions of 'truths cotenable with the antecedent' and 'background propositions'. A possible way to defend UI might be to deny that $A \to C$ and $A \to \neg C$ can ever both be false. However, that would not be a correct interpretation of Goodman either, since he said that $A \to C$ and $A \to \neg C$ are contraries and can both be false[18].

TD requires that $\{A\} \cup S$ be self-compatible (and self-cotenable as Goodman adds later). This assumes that A by itself must be self-compatible, i.e. possible. Goodman wanted to consider 'counterlegals' (counterfactuals with impossible antecedents) separately. UI is more general. Omitting the requirement for self-compatibility, UI applies to counterlegals as well.

---

[18] Cf. ibid. p. 18 footnote 8.

Thus we can say that there is a reasonable justification for each of the differences that we considered so far between UI and Goodman's theory, except in the case of the condition from Goodman's footnote 6. What is left to be commented on is Goodman requirement that S be compatible (cotenable?) with both C and ¬C. Let me skip this for now. I will comment on that later. My topic is not Goodman's theory by itself, but rather the relation between Goodman's theory and various possible worlds semantics for counterfactuals. I will continue my discussion of Goodman's theory after introducing some possible worlds semantics, so that I could compare them.

3. Stalnaker

Now, 2.2 is true and 2.3 is false, but why is it so? A possible explanation might go this way. Imagine that the world suddenly changed only insofar as it were necessary for the match m to be struck. Would in that case the match m (a) light, or (b) be wet? Why would it light? Because dry etc matches light when struck – this is a law not subjected to the suggested necessary alterations. Hence there is no need to postulate any further changes to explain (a) other than those necessary ones. Why would it be wet? Sellars[19] gives an example in which Tom asks Dick why does he claim 2.3 and Dick explains: "Well, Harry is over there, and he has a phobia about matches. If he sees anyone scratch a match, he puts it in water." But in our example Harry is not around, so an explanation for (b) would require some further changes beside the mentioned necessary ones, because striking does not wet, nor does the match need to be wet in order to be struck. The criterion for truth of counterfactual conditionals would thus be whether for the explanation of the consequent being true there is need to postulate any changes other than the necessary ones for the antecedent to be true. That way Goodman's problem of choosing between 2.2 and 2.3 seems to be *solved*, while the problem of cotenability is *avoided*[20]. Let me emphasize this idea:

3.0 $A \rightarrow C$ is true iff C will be true after only the changes necessary to make A true have been made.

---

[19] Sellars 1957

[20] A somewhat similar explanation can be found in Sosa (ed.) 1975, Introduction pp. 13-14.

The previous paragraph and 3.0 are usually taken as presenting the basic idea of the *minimal change* theory of counterfactuals. The first such theory was proposed by Stalnaker[21]. A→C is true according to his theory iff C holds at *the most similar* A-world, i.e. at the world where A is true and which is more similar to the actual world than any other A-world. Thus the idea of minimal change in our world is captured by the notion of similarity, which remains undefined. On that point we see that Stalnaker's project is less ambitious than Goodman's. Stalnaker allows a vague notion (similarity) in his semantics, as long as that notion enables him to do some formal work. It enables him to give a formal semantics, axioms and rules of inference, and to prove completeness. Here I will present a formal system made by Nute[22], but equivalent to the system C2 Stalnaker proposed together with Thomason[23]. Nute's axiomatization makes it easier to compare C2 with other conditional logics.

A model for Nute's version of Stalnaker's logic is a quadruple < I, R, s, [] > where I is a set of possible worlds, R a binary reflexive (accessibility) relation on I, s a partial world selection function, which, when defined, assigns to proposition ϕ and a world $i \in I$ a world $s(\phi, i)$ (the ϕ-world the most similar to i), and [] is a function which assigns to each proposition ϕ a subset $[\phi]$ of I (all the worlds from I where ϕ is true). Thus Stalnaker's conditional logic, like many conditional logics that appeared later, consists of modal logic with the addition of one primitive term – the selection function s. The intuitive sense of that function is to capture the meaning of similarity of possible worlds. Since similarity is a very vague notion, we cannot hope to define it, formally or informally. But we can express formally *some* of the features of that notion. And these formal expressions will determine the formal properties of the arrow '→'. S1-S5 are the properties of the selection function in Stalnaker's system, and S6 is the definition of truth conditions for '→':

---

[21] Stalnaker 1968.

[22] Nute and Cross 2002, pp. 9-10.

[23] Cf. Stalnaker and Thomason 1970. For subtle differences, which do not need to concern us here, between C2 and Nute's system see Nute pp. 10-11.

S1 $s(\phi, i) \in [\phi]$

(ϕ holds at the ϕ-world most similar to i)

S2 $<i, s(\phi,i)> \in R$

(the ϕ-world most similar to i is accessible from i)

S3 if $s(\phi, i)$ is not defined then for all $j \in I$ such that $<i, j> \in R$, $j \notin [\phi]$

(if the ϕ-world most similar to i is not defined then ϕ is impossible at i)

S4 if $i \in [\phi]$ then $s(\phi, i) = i$

(if ϕ is true at i, than i is the ϕ-world most similar to i)

S5 if $s(\phi, i) \in [\psi]$ and $s(\psi, i) \in [\phi]$, then $s(\phi, i) = s(\psi, i)$

(if ψ holds at the ϕ-world most similar to i, and ϕ holds at the ψ-world most similar to i, then the ϕ- and the ψ-worlds most similar to i are the same)

S6 $i \in [\phi \to \psi]$ iff $s(\phi, i) \in [\psi]$, or $s(\phi, i)$ is undefined

(a counterfactual conditional $\phi \to \psi$ is true at the world i iff ψ holds at the ϕ-world most similar to i)

Stalnaker's semantics can be axiomatized by adding to the classical propositional logic the following two rules and six axioms:

RCEC
$$\frac{\phi \equiv \psi}{(\chi \to \phi) \equiv (\chi \to \psi)}$$

RCK	$\dfrac{(\phi_1 \wedge \ldots \wedge \phi_n) \supset \psi}{(\,(\chi\to\phi_1) \wedge \ldots \wedge (\chi\to\phi_n)\,) \supset (\chi\to\psi)}$	$(n \geq 0)$
ID	$\phi\to\phi$	
MP	$(\phi\to\psi) \supset (\phi\supset\psi)$	
MOD	$(\neg\phi\to\phi) \supset (\psi\to\phi)$	
CSO	$(\,(\phi\to\psi) \wedge (\psi\to\phi)\,) \supset (\,(\phi\to\chi) \equiv (\psi\to\chi)\,)$	
CV	$(\,(\phi\to\psi) \wedge \neg(\phi\to\neg\chi)\,) \supset (\,(\phi\wedge\chi)\to\psi\,)$	
CEM	$(\phi\to\psi) \vee (\phi\to\neg\psi)$	

The last one, CEM, is the distinctive feature of Stalnaker's conditional logic. CEM is an abbreviation for 'conditional excluded middle'. CEM is highly controversial. Despite Stalnaker's defense[24], there seem to be more philosophers who reject it than those who accept it.

I am now interested in the relation between Stalnaker's and Goodman's theory. It was noticed early that Goodman's theory also has something to do with CEM. Pollock proves that Goodman's theory validates CEM for conditionals with false antecedents, and uses that as an argument against Goodman:

> Given this difference [between $\neg(P\to\neg Q)$ and $P\to Q$, which are not equivalent for Pollock because he rejects CEM] I think it is clear that Goodman's requirement of cotenability is too weak. This is demonstrated by seeing that it would lead right back to a special case of the principle that if $\ulcorner\neg(P\to\neg Q)\urcorner$ is true then $\ulcorner P\to Q\urcorner$ is true. More precisely, Goodman's proposal implies that whenever P is false and

---

[24] Stalnaker 1978.

> $\ulcorner \neg(P \rightarrow \neg Q) \urcorner$ is true, then $\ulcorner P \rightarrow Q \urcorner$ is true. This implication is established as follows. First, we need two obvious principles regarding subjunctive conditionals:
>
> (1) If $\ulcorner P \rightarrow Q \urcorner$ is true and Q entails R, then $\ulcorner P \rightarrow R \urcorner$ is true.
>
> (2) If $\ulcorner P \rightarrow (P \supset Q) \urcorner$ is true, then $\ulcorner P \rightarrow Q \urcorner$ is true.
>
> (1) is so obvious as to need no defence. (2) holds because if $\ulcorner P \supset Q \urcorner$ would be true if P were true, then both P and $\ulcorner P \supset Q \urcorner$ would be true if P were true, and hence Q would have to be true if P were true. Given these principles, let us suppose, with Goodman, that truth and cotenability are all that is required for inclusion in C [Pollock uses 'C' for the set of background propositions that Goodman called S]. Suppose P is false and $\ulcorner \neg(P \rightarrow \neg Q) \urcorner$ is true. Then by (1) $\ulcorner \neg(P \rightarrow (P \wedge \neg Q)) \urcorner$ is true and so $\ulcorner \neg(P \rightarrow \neg(P \supset Q)) \urcorner$ is true. But as P is false, $\ulcorner P \supset Q \urcorner$ is true, and if follows from Goodman's proposal that $\ulcorner P \rightarrow (P \supset Q) \urcorner$ is true. Then from (2) it follows that $\ulcorner P \rightarrow Q \urcorner$ is true.[25]

The proof became pretty fast at the end, so I will repeat it here. Assume

3.1 P is false, and

3.2 $\neg(P \rightarrow \neg Q)$ is true.

Then note that by contraposition the principle marked by (1) in the citation is equivalent to:

3.3 If $P \rightarrow R$ is false then either $P \rightarrow Q$ is false or Q does not entail R,

which is equivalent to:

3.4 If $P \rightarrow R$ is false and Q entails R, then $P \rightarrow Q$ is false.

---

[25] Pollock 1976 p. 11.

The rest of the proof goes as follows

3.5 $A\wedge\neg Q$ entails $\neg Q$

3.6 $\neg(P\rightarrow(P\wedge\neg Q))$ from 3.2 and 3.5 by 3.4 and substitution of $P\wedge\neg Q$ for Q and $\neg Q$ for R

3.7 $P\wedge\neg Q$ is equivalent to $\neg(P\supset Q)$

3.8 $\neg(P\rightarrow\neg(P\supset Q))$ from 3.6 and 3.7. This means by Goodman's definition that $P\supset Q$ is cotenable with P.

3.9 $P\supset Q$ is true from 3.1

3.10 $P\rightarrow(P\supset Q)$ from 3.8 and 3.9 by what Pollock called 'Goodman's proposal', i.e. a counterfactual is true iff the consequent is entailed by the antecedent and the truths cotenable with the antecedent (which I called UI in the previous chapter).

3.11 $P\rightarrow Q$ from 3.10 by the principle (2) from the citation.

QED

From the citation we can see that Pollock reads Goodman the same way as Nute does ("truth and cotenability are all that is required for inclusion in C"), so Pollock's conclusion holds for Nute's version of the usual interpretation of Goodman (UI). Note that the principles (1) and (2) are 'safe', that is, they can easily be derived from UI.

Bennett would agree with Pollock that it is enough for a theory to imply CEM to be rejected, but he came closer to a more general result. It seems that he believes, or at least I find it plausible to interpret Bennett as if he believes that Nute's version of UI implies CEM without any restriction. He did not say so explicitly, but he said that

Goodman's theory (which he reads the same way as Nute does) was "worryingly reminiscent" of something else, which Bennett calls PF, and which does imply CEM[26]. PF is a criterion for similarity between possible worlds that Bennett considers in another context and rejects[27]. We do not need to consider PF here. It is enough to know that PF implies PF*[28]:

PF* $C \wedge \neg(A \rightarrow \neg C)$ entails $A \rightarrow C$

In yet another context Cross proved that PF* entails CEM[29]. This can be proven as follows.

Theorem 1 If PF* is valid, so is CEM.

Proof: Suppose the negation of CEM

$$\neg(A \rightarrow C) \wedge \neg(A \rightarrow \neg C)$$

Then suppose C. Together with the right conjunct $\neg(A \rightarrow \neg C)$, C implies $A \rightarrow C$ by PF*. But this contradicts the left conjunct $\neg(A \rightarrow C)$. Now suppose $\neg C$. Given that $A \rightarrow C$ is equivalent to $A \rightarrow \neg\neg C$, $\neg C$ and the left conjunct $\neg(A \rightarrow C)$ by PF* imply $A \rightarrow \neg C$, because another instance of PF* is: $\neg C \wedge \neg(A \rightarrow \neg\neg C)$ entails $A \rightarrow \neg C$. But $A \rightarrow \neg C$ contradicts the right conjunct. Therefore CEM.

QED

---

[26] Cf. Bennett 2003 p. 309.
[27] Ibid. p.232.
[28] Ibid. p. 233.
[29] Cross 1985.

Therefore, if PF* is valid, so is CEM. I use this proof because it is shorter and good enough for my present purposes. Cross's proof is longer, but he uses only classical logic and makes no assumptions about conditionals, while I supposed that equivalent consequents can be substituted. This proof shows that Nute's UI validates CEM, because PF* is obviously implied by Nute's UI: PF* says that A→C is true whenever C is true and cotenable with A, and A→C is true according to UI iff A and all cotenable truths entail C, i.e. $\{A\}\cup S \vdash C$. If C is true and cotenable with A, then it is already in S, so, trivially, it is entailed by $\{A\}\cup S$, and A→C follows.

There is an easier way to prove that CEM is validated by Nute's UI, a way which also offers an intuitive explanation of what is going on. It is enough to note that any possible proposition plus every truth cotenable with it are enough to determine a single world. That is, the antecedent A and the set S of background propositions as described by Nute' UI determine uniquely one world. That means that $\{A\}\cup S$ is a subset of only one maximal possible set. In the world where all the members from $\{A\}\cup S$ hold, either C is true or ¬C is. Thus either A→C or A→¬C must be true, because either $\{A\}\cup S \vdash C$ or $\{A\}\cup S \vdash \neg C$ must hold. So CEM follows from Nute's UI.

Theorem 2 $\{A\}\cup S$ determines uniquely one world.

Proof: Suppose there are two different worlds *j* and *k* such that every proposition from $\{A\}\cup S$ holds at both *j* and *k*. S is a set of truths from the actual world *i* that are cotenable with A. Since $j\neq k$ there is a proposition D true at the actual world such that $j\models D$ and $k\models\neg D$.

D is either cotenable with A or not. If not, then $i \models A \to \neg D$, which means that $\{A\} \cup S \vdash \neg D$, and j is then an impossible world. If D is cotenable with A, then D is in S and then k is an impossible world. Therefore $j = k$.

QED

Thus the theorems 1 and 2 give us the syntactic and the semantic version of the same claim.

Two things should be noted again (both discussed in the previous chapter). First, Goodman rejects CEM. He said that $A \to C$ and $A \to \neg C$ are contraries and can both be false[30]. Second, from Goodman's paper we cannot tell whether Nute's UI (which validates CEM) or Lewis's UI (which does not validate CEM, since it says that S is finite) is a correct interpretation.

Why do so many people interpret Goodman as if he said that S includes every cotenable truth? I guess because it appears to be easier to rule out propositions that are in a kind of collision with the antecedent than to define the set of relevant propositions. Besides that, in monotonic logic – and classical logic on which Goodman bases the validity of the argument 2.1 is monotonic – a valid argument remains valid if we add to it more premises, whether they are relevant for the conclusion or not. Bennett explains (italics are mine):

---

[30] Cf. Goodman 1991. p. 18 footnote 8.

> Irrelevant conjuncts are mere clutter: *they cannot lead to any conditional's being accorded a truth value that it does not deserve*. [Possible] Worlds analysis [of counterfactuals] take in vast amounts of irrelevant materials, and clearly get away with it. A Worlds theorist will say that the truth value of 'If you had unplugged the computer, it would not have been damaged by lightning' depends upon what obtains at certain worlds that are just like α up to a certain moment... *Just* like α? Worlds resembling α in respect of the number of sardines in the Atlantic, the average colour of alpine lilies in Tibet, and the salinity of the smallest rock pool in Iceland? What have those to do with the conditional about the computer? Nothing, but Worlds theories bring them in because they are too much trouble to keep out, and – the main point – they do no harm. *Irrelevance is harmless.*[31]

However, we saw that irrelevance has influence (whether you want to call that influence harmful or good is a further question). For those who reject CEM, irrelevant conjuncts do lead to a conditional being accorded a truth value that it does not deserve, namely they force us to consider one of A→C and A→¬C true even when we think that both are false. Therefore, the question whether S contains irrelevant propositions (and if so, how many) is not trivial, because *the answer has influence on the formal properties of '→'*.

It should be noted that although Nute's UI validates CEM, it is still not equivalent to Stalnaker's theory. This is because Nute's UI does not validate S5 nor its syntactic counterpart CSO. Here is a countermodel to S5 and CSO in which the truth conditions for '$\rightarrow$' are defined as in Nute's UI. Suppose that at the actual world α the following hold:

---

[31] Bennett 2003 pp. 307-8.

3.12 $\alpha \models \phi \rightarrow \psi$

3.13 $\alpha \models \psi \rightarrow \phi$

3.14 $\alpha \models \phi \rightarrow \chi$

but

3.15 $\alpha \not\models \psi \rightarrow \chi$

3.12, 3.13, and 3.14 respectively mean that there are valid arguments

3.16 $\{\phi\} \cup B_{\phi} \vdash \psi$

3.17 $\{\psi\} \cup B_{\psi} \vdash \phi$

3.18 $\{\phi\} \cup B_{\phi} \vdash \chi$

where B_{ϕ} is the set of all truths from α cotenable with ϕ, and similarly for B_{ψ}. We know that $\{\psi\}$ and B_{ψ} uniquely determine a world, or a maximal consistent set. From 3.15 we know that $\{\psi\} \cup B_{\psi}$ does not entail χ. So it must entail $\neg\chi$, because a maximal set must contain $\neg\chi$ if it does not contain χ. Therefore (as CEM also implies) from 3.15 it follows

3.19 $\{\psi\} \cup B_{\psi} \vdash \neg\chi$

and

3.20 $\alpha \models \psi \rightarrow \neg\chi$

Suppose also:

3.21 $\alpha \models \neg\phi \wedge \neg\psi \wedge (\phi \supset \psi) \wedge (\psi \supset \phi) \wedge (\phi \supset \chi) \wedge (\psi \supset \neg\chi)$

and that

3.22 $(\phi \supset \psi) \wedge (\phi \supset \chi)$

is cotenable with ϕ, and

3.23 $(\psi \supset \phi) \wedge (\psi \supset \neg\chi)$

is cotenable with ψ. Since 3.22 and 3.23 hold at α (i.e. they are true), they are contained in B_ϕ and B_ψ respectively.

3.12 and 3.14 are then true because ϕ and 3.22 entail ψ and χ, and 3.13 and 3.20 are true because ψ and 3.23 entail ϕ and $\neg\chi$. 3.15 follows from 3.20. The world determined by $\{\phi\}$ and B_ϕ (in Stalnaker's notation – $s(\alpha,\phi)$) is described by the maximal consistent set based on ϕ and 3.22 and other truths from α cotenable with ϕ, while the world $s(\alpha,\psi)$ determined by $\{\psi\}$ and B_ψ is described by the maximal consistent set based on ψ and 3.23, and other truths from α cotenable with ψ. Therefore, $s(\alpha,\phi) \neq s(\alpha,\psi)$, since $s(\alpha,\phi) \models \chi$ and $s(\alpha,\psi) \models \neg\chi$. This completes the countermodel.

Thus CSO is valid in Stalnaker's semantics and invalid in Nute's UI. Which theory scores a point here? It is not easy to answer. I will comment on CSO again in chapters 6 and 7 (sections 6.3 and 7.3) and we will see that lots of other issues about counterfactuals are involved.

Conditional excluded middle, as we saw, is the distinctive feature of Stalnaker's formal system, and the semantic counterpart of that feature is what David Lewis calls 'Stalnaker's assumption', according to which there is always a unique most similar world for any possible antecedent. However, thinking in terms of our understanding of the word 'similarity' and our metaphysical intuitions about possible worlds, it seems that there are cases where it is not possible to decide which of two (or more) worlds is more similar.

There are cases where the most similar world does not even seem to exist, for example the most similar world where this line

is shorter than it actually is[32]. Lewis's semantics captures these intuitions against Stalnaker's assumption. A counterfactual $A \rightarrow C$ is true according to Lewis iff C holds in *all* most similar A-worlds, or, more generally, if there are no most similar A-worlds as it is the case in the line example, then $A \rightarrow C$ is true iff there are $A \wedge C$-worlds more similar than any $A \wedge \neg C$-world. If A is impossible, $A \rightarrow C$ is 'vacuously' true.

Of course, our metaphysical intuitions about possible worlds are not a reason to reject CEM. Why does Lewis propose this semantics that opposes Stalnaker's assumption and makes CEM invalid? He offers two arguments against CEM. The first relies on an intuition that both 3.24 and 3.25 are false[33]

3.24 If Bizet and Verdi were compatriots, they would be Italians

3.25 If Bizet and Verdi were compatriots, they would not be Italians (but French)

That is, the first argument relies on an intuition that there are cases where a proposition A does not counterfactually imply either C or $\neg C$.

The second argument is that CEM annihilates the distinction, which Lewis finds intuitively very appealing, between 'would' and 'might' conditionals[34]. The 'would' conditionals are the counterfactuals we have been representing so far with the arrow '$\rightarrow$'. The 'might' conditionals are defined in terms of 'would' as $A \Diamond\!\!\rightarrow C \stackrel{\text{def}}{=} \neg(A \rightarrow \neg C)$.

---

[32] This example is from Lewis 1973 p. 20.
[33] Cf. ibid. p. 80.
[34] Cf. ibid.

'A◊→C' is read 'If it were the case that A, it might have been the case that C'[35]. Given the definition above, A◊→C is true iff A is possible and there is no A∧¬C-world more similar than any A∧C-world[36] (which is the case when A→¬C is false). Whenever the antecedent A is possible, A→C implies A◊→C. Assuming CEM, the converse implication from 'might' to 'would' also holds, and that makes the two conditionals equivalent whenever the antecedent is possible.

Proof (assuming that A is possible):

3.26 $(A \to C) \supset \neg(A \to \neg C)$

3.26 is a principle that follows directly from Lewis's truth conditions for '→' (if C holds at the closest A-worlds, then ¬C does not). The consequent of 3.26 is by definition equivalent to A◊→C, so

3.27 $(A \to C) \supset (A \Diamond\!\!\to C)$

3.28 $(A \to C) \vee (A \to \neg C)$ CEM

3.29 $\neg(A \to \neg C) \supset (A \to C)$ from 3.28 by classical logic

3.30 $(A \Diamond\!\!\to C) \supset (A \to C)$ from 3.29 by definition of '◊→'

3.26 and 3.30 establish the equivalence.

QED.

Wanting to preserve the might-would distinction, Lewis rejects CEM. This is the main difference between Lewis's and Stalnaker's theories, which are otherwise very similar in many respects.

---

[35] Cf. ibid. p. 21.
[36] Ibid.

Interesting papers have been written to defend CEM from Lewis's attack[37]. I will not comment on them here, because my reasons for rejecting CEM are quire different from Lewis's. My logical taste is on the side of relevance logic. I believe therefore that A→C and A→¬C can both be false when A has nothing to do with C and ¬C. There has to be a relevant connection between the antecedent and the consequence for a conditional to be true. This leads me to a disagreement with both Stalnaker and Lewis:

3.31 If one were to scare a pregnant guinea pig, then all her babies would be born without tails.

3.32 If one were to scare a pregnant guinea pig, then some of her babies would not be born without tails.[38]

CEM predicts that one of these would be true, while they are both false in relevance logic. 3.31 is true according to Lewis because the consequent's being false represents a relatively big departure from actuality compared to the antecedent's being true. Since the antecedent has nothing to do with the consequent, the consequent will still hold at the closest antecedent-worlds. In general, many counterfactuals of the form (small change) → (not big change) are true for Lewis. Lewis's might-conditionals, however, commit even more fallacies of relevance. Whenever a would-conditional is false, that is, whenever Lewis's semantics denies the would-connection between an antecedent and a consequent, it asserts the might-connection between the antecedent and the negation of the consequent, although the antecedent might be irrelevant for both consequent and its negation. For example, 3.32 is false in Lewis's semantics (guinea pigs have no tails, their mother being frightened or not), which implies the might conditional

---

[37] Especially Stalnaker 1978, van Fraassen 1966.

[38] This is a modification of an example that Michael Dunn ascribes to Alan Ross Anderson, cf. Dunn 1986.

P

3.33 If one were to scare a pregnant guinea pig, then it might be that all her babies would be born without tails.

If 3.33 appears to be true, I think this is only because of confusing 3.33 with the corresponding even-if conditional

3.34 Even if one were to scare a pregnant guinea pig, it might be that all her babies would be born without tails.

3.34 denies, while 3.33 asserts that certain connection obtains between scaring and having tails, which makes 3.34 true and 3.33 false.

Now I will switch my attention from CEM back to cotenability and the relation between the metalinguistic and possible worlds theories of counterfactuals.

4. Lewis

The crucial notion of Lewis's theory of counterfactuals is similarity. The crucial notion of the metalinguistic theory is cotenability. Lewis defined cotenability in terms of similarity and explained what he thought to be a connection between his and the metalinguistic theory of counterfactuals. The purpose of this chapter is to deny the existence of that connection and to give a counterexample to Lewis's notion of cotenability.

First I will remind the reader of Lewis's theory and terminology. In his book *Counterfactuals*, Lewis states his definition of the truth conditions for counterfactuals in several different ways. I will discuss five of them: defining the truth conditions in terms of systems of spheres, in terms of comparative 'overall' similarity, in terms of comparative possibility, in terms of inner necessity, and in terms of cotenability. The intuitive sense of all these notions can be expressed in terms of comparative overall similarity. For each possible world i and a set S_i of worlds accessible from i, and for any $j, k \in S_i$ the relation of similarity '$\leq$' tells us whether the world j is at least as similar to the world i as the world k is ($j \leq_i k$), or the other way around ($k \leq_i j$), or both (if j and k are equally similar to i then both $j \leq_i k$ and $k \leq_i j$ are true). This relation is called '*overall*' similarity probably because, as Lewis explained, it "consists of innumerable similarities and differences in innumerable respects of comparison, balanced against each other according to the relative importances we attach to those respects of comparison."[39]. It is also called 'absolute' similarity in the literature, which means that it is not in any way relative to the conditional or conditionals that are being evaluated (there are other theories for counterfactuals that use a notion of similarity or selection function that is relative either to the antecedent, or both antecedent and consequent, or context; these will be discussed in chapter 6). In terms of similarity, a counterfactual conditional $\phi \rightarrow \psi$ is true at a world *i* iff

---

[39] Lewis, 1973 p. 91.

(Definition 1 – comparative similarity)

either ϕ is impossible, that is, no ϕ-world (i.e. a world at which ϕ holds) is accessible from *i*,

or there is a world $j \in S_i$ at which both ϕ and ψ hold and *j* is more similar to *i* than any world at which both ϕ and $\neg\psi$ hold.[40]

A system of spheres \$ is an assignment to each possible world *i* of a set $\$_i$ of nested sets of possible worlds accessible from *i*.[41] In a picture we can represent the set $\cup\$_i$ of all the worlds accessible from *i* as a sphere, the world *i* as the centre of that sphere, and each possible world from $\cup\$_i$ as a point in the sphere. Inside the sphere we can draw more concentric spheres with *i* as a centre. Each such sphere represents one member of $\$_i$. A system of spheres carries information about the similarity of worlds: whenever a world belongs to a sphere and another world is outside that sphere, the first world resembles *i* more closely than the second world does. $\phi \rightarrow \psi$ is true at *i* according to a system of spheres \$ iff

(Definition 2 – systems of spheres)

either ϕ is impossible,

---

[40] Cf. ibid. p. 49.

[41] Cf. ibid. pp. 13-14.

or (as on the figure 1) some sphere $S \in \$_i$ contains some ϕ-worlds and all of these ϕ-worlds are also ψ-worlds.[42]

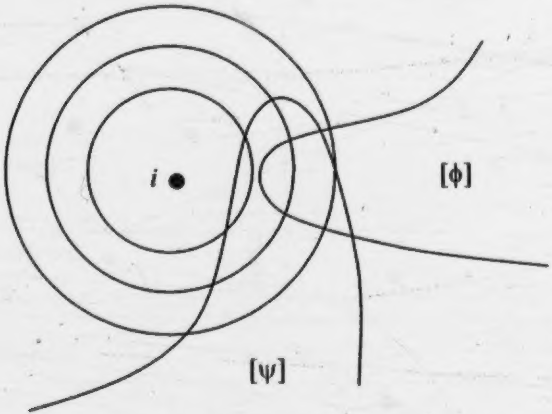

Fig. 1.

$[\phi]$ and $[\psi]$ stand for sets of worlds at which ϕ and ψ (respectively) hold. Lewis defines a proposition as a set of worlds at which the proposition holds. He nevertheless uses different symbols for the definiens and the definiendum: 'ϕ' stands for the proposition ϕ and '$[\phi]$' stands for the set of all and only ϕ-worlds.

A proposition ϕ is said to be *more possible* than ψ (relative to a world *i*) iff there are ϕ-worlds closer to *i* than any ψ-world (i.e. if some sphere contains some ϕ-worlds but no ψ-world)[43]. The same thing expressed in Lewis's symbols is $\phi \prec_i \psi$ or $[\phi] \prec_i [\psi]$. Thus $\phi \rightarrow \psi$ is true at *i* iff:

(Definition 3 – comparative possibility)

either ϕ is impossible,

or $\phi \wedge \psi \prec_i \phi \wedge \neg\psi$

---

[42] Cf. ibid. p. 16.
[43] Ibid. p. 52.

The truth conditions for '$\rightarrow$' can be reformulated in terms of *inner* necessity. Lewis calls a proposition ϕ inner necessary (relative to a world *i*) iff there is a sphere $S \in \$_i$ such that ϕ holds at every world from S.[44] 'Outer' necessity is the usual notion of necessity, viz. truth at every accessible world. Intuitively, the negation of an inner necessary proposition represents a relatively big departure from actuality; the bigger the scope of necessity, i.e. the bigger the sphere throughout which a proposition holds, the bigger the departure. $\phi \rightarrow \psi$ is true iff

(Definition 4 – inner necessity)

either ϕ is impossible

or the material conditional $\phi \supset \psi$ is inner necessary and the scope of its necessity is big enough to include some ϕ-worlds.

From this definition we can see why Lewis calls his counterfactuals *variably strict conditionals*[45]. C. I. Lewis's strict implication is equivalent to a necessary material implication, that is, a material implication true at all accessible worlds. David Lewis's counterfactual is stronger than material implication and weaker than strong implication. Within a subset of all accessible worlds (i.e. within a sphere), David Lewis's counterfactual behaves like a strong implication. That subset (sphere) varies from counterfactual to counterfactual, and hence the name 'variably strict' conditional.

All the definitions of truth conditions stated above depend on a notion that can be defined in terms of similarity. Thus Lewis's conditional logic consists of classical propositional logic, plus modal logic, plus one more primitive term – similarity relation.

---

[44] Ibid. p. 30.

[45] Cf. ibid. p. 13.

This makes Lewis's logic very similar to Stalnaker's, but there is a philosophically important difference that is nicely summarized by Dorothy Edgington:

> Between Stalnaker and Lewis, there are differences in formulation, and some substantive differences, but also a difference in aim. Stalnaker's project is less ambitious. He does not expect there to be an informative analysis of "A-world which differs minimally from the actual world" which could be specified independently of judgments about what would have been true if A were true. Lewis seeks a genuine analysis of counterfactuals in terms which do not presuppose them.[46]

If this is correct, than Goodman's, Stalnaker's and Lewis's ambitions can be summarized as follows. Goodman attempted a genuine (i.e. reductive) analysis of counterfactuals in terms of well-defined and precise notions of logic (of his time). Stalnaker gives up the reductive analysis and ultimately accept circularity. He does not think, as Goodman did, that circularity is a failure, because similarity is a convenient notion that enables us to make logical systems. That is also a justification for using that notion although it is not well defined, but vague. Lewis accepts vagueness, but attempts a reductive analysis, without circularity.

If both Goodman's and Lewis's theory are reductive, and if Lewis's theory is successful, then on the basis of Lewis's theory we should be able either to prove that Goodman was wrong and his theory not worth dealing with, or to finish his project and solve the problem he could not solve – to define cotenability without circularity or infinite regress. Lewis does not think that Goodman was wrong. He thinks that the intuitions behind the two theories are in accordance and that using Lewis's superior terminology they can be shown to be equivalent. Here is Lewis's definition of cotenability, which I will call cotenability$_l$ in order to distinguish it from Goodman's notion cotenability$_g$:

---

[46] Edgington 1995. p. 251.

> "Let us say that χ is cotenable with ϕ at a world *i* (according to a system of spheres \$) iff either (1) χ holds throughout $\cup\$_i$, or (2) χ holds throughout some ϕ-permitting sphere in $\$_i$. In other words: iff either (1) χ holds at all worlds accessible from *i*, or (2) some ϕ-world is closer to *i* than any $\neg\chi$-world. A necessary truth (in the sense of outer necessity) is cotenable with anything; a falsehood is cotenable with nothing. Between these limits, cotenability is a matter of comparative possibility. If ϕ is entertainable at *i*, χ is cotenable with ϕ at *i* iff $[\phi]<_i[\neg\chi]$."[47]

Using the notion of cotenability$_l$ Lewis can express his definition of truth conditions for '$\rightarrow$' with the recognizable metalinguistic pattern: $\phi\rightarrow\psi$ is true at *i* iff

(Definition 5 – cotenability)

there are finitely many propositions $\chi_1,\ldots\chi_n$, each cotenable$_l$ with ϕ at *i*, such that the following argument is valid:

$$\frac{\phi, \chi_1,\ldots,\chi_n}{\psi}$$

---

[47] Ibid. p. 57. "Entertainable" in this context means "possible". $[\phi]<_i[\neg\chi]$ means that there are ϕ-worlds that are more similar to *i* than any $\neg\chi$-world is. Note that according to this definition a proposition must be inner necessary if it is cotenable with something.

Replacing cotenability$_l$ with cotenability$_g$ in Definition 5 gives us what I have called Lewis's version of the usual interpretation (UI) of Goodman's theory (chapter 2).

Cotenability$_l$ is a notion significantly different from cotenability$_g$. (As we said in the chapter 2, A is cotenable$_g$ with B iff $B \rightarrow \neg A$ is false). A false proposition can be cotenable$_g$ with another proposition, but no falsity is cotenable$_l$ with anything. This is one of the reasons why Lewis's notion should not be understood relative to its etymology or to some possible natural language meaning of 'cotenable' (co-tenable, jointly tenable, possibly true at the same time). Cotenability$_l$ should be understood as an artificial notion (which, by itself, says nothing against Lewis's notion, but is important for its understanding). The purpose of the two notions, as we can conclude based on the comments from chapter 2, is different. Cotenability$_g$ is only one condition, necessary but not sufficient, for a proposition to be included among the background propositions (cf. pp 19-20 above). On the other hand, any proposition cotenable$_l$ with an antecedent can be a background proposition. Another difference is that the modal status of any proposition cotenable$_l$ with a falsity is not simply truth, but inner or outer necessity. Nothing similar can be inferred from cotenability$_g$. Note also that cotenability$_l$ implies cotenability$_g$, but not the other way around.

Despite these differences, Lewis claims that cotenability$_l$ is the link between the two theories. Here is how he explains the relation between them:

> On any metalinguistic theory, the principal problem is to specify which further premises $\chi_1,\ldots,\chi_n$ are suitable to be used with a given antecedent and which are not. The metalinguistic theorist uses his further premises in much the same way as

> I have used the system of spheres representing comparative similarity of worlds: to rule out of consideration many of the various ways the antecedent could hold, especially the more bizarre ways. $\phi \rightarrow \psi$ is true or assertable on a metalinguistic theory iff ψ holds at all ϕ-worlds of a certain sort: ϕ-worlds at which some further premises, suitable for use with the antecedent ϕ, hold.[48]

Let us use Goodman's example with the match m to make Lewis's idea clear. Let ϕ = I strike the match m; ψ = m lights; χ_1 = m is dry; χ_2 = m is well made; χ_3 = oxygen enough is present; χ_4 = All dry, well made matches light when struck in the presence of oxygen; and let χ without a subscript be the conjunction of the latter four propositions. Let the situation be as before: χ is true and ϕ and ψ are false at the actual world *i*. Obviously, the following conditional is true at *i*:

4.1 Had I struck the match m, it would have lit ($\phi \rightarrow \psi$)

We evaluate 4.1 in Lewis's semantics by looking at the relevant ϕ-worlds to see if ψ holds there. What are the relevant ϕ-worlds? Those at which χ holds. This makes perfect sense. If we need to know the truth value of 4.1, we are interested in the worlds where m is struck, dry, well made, oxygen is present and the same laws of nature hold as in the actual world. We are obviously not interested in the worlds where the same laws hold, but the match is wet, or in the worlds where matches burn when put into water. Thus we have a connection between Lewis's and the metalinguistic theory: the most similar worlds important for Lewis are those where the background propositions from the metalinguistic theory hold. This simple and attractive idea is, I think, what governed Lewis in his description of the relation between the two theories. Besides that, this idea is very helpful in understanding what Lewis meant by the notion of similarity. That notion is notoriously vague and lots of people complain that from Lewis's writings they cannot understand what exactly it is supposed to mean. The passage cited above certainly does not give the final answer to the question of meaning of similarity. Nevertheless, it is helpful because it gives us some priorities in estimating how close or remote worlds are to our world: we should pay more attention to whether the background propositions hold at those worlds or not.

---

[48] ibid p. 66

Still it seems that Lewis's theory says a bit more than the metalinguistic theory. χ does not only hold at the closest ϕ-worlds. It holds throughout a ϕ-permitting sphere. On Figure 2 χ holds throughout the sphere S. Goodman's metalinguistic theory cannot express this strange modality of background propositions nor can it offer any reason why background propositions should have any modality stronger than truth. Lewis, however, thinks that the inner necessity of background propositions is in accordance with the metalinguistic theory:

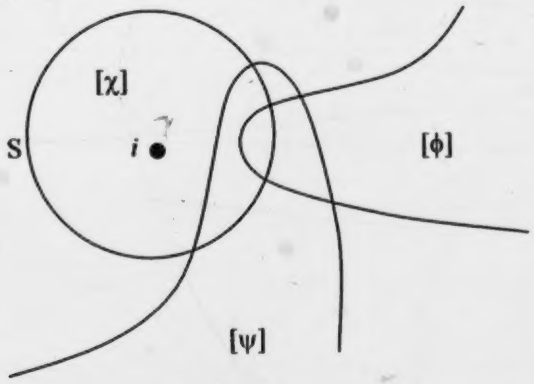

Fig. 2.

> I think that my definition of cotenability ... captures the intentions of metalinguistic theorists. On my definition, a cotenable premise is not only true, but also necessary to some extent. The strictness of its necessity is the least strictness that will not rule the antecedent out as impossible, provided that the antecedent is entertainable so that some such strictness exist."[49]

Unfortunately, Lewis did not explain what intentions of metalinguistic theorists he had in mind[50]. In any case, as we will see, inner necessity is a big part of the problem that will be discussed shortly.

It is easy to see that definitions 1 – 4 are equivalent[51]. It is less obvious that Definition 5 is equivalent to the other four, so I cite Lewis's proof:

> A counterfactual $\phi \rightarrow \psi$ is true at *i* (according to my truth conditions) iff the premise ϕ and some auxiliary premise χ cotenable with ϕ at *i*, logically imply ψ. Proof: Suppose there is some such premise χ. Perhaps there is no ϕ-permitting sphere around *i*, in which case $\phi \rightarrow \psi$ is vacuously true at *i*. Otherwise there is a ϕ-permitting sphere throughout which χ holds; since ϕ and χ jointly imply ψ, $\phi \supset \psi$ also holds throughout sphere; so $\phi \rightarrow \psi$ is true. Conversely, suppose $\phi \rightarrow \psi$ is true at *i*. Either there is no ϕ-permitting sphere around *i*, in which case $\neg\phi$ is a premise, cotenable with ϕ at *i*, which together with ϕ implies ψ; or else there is a ϕ-

---

[49] ibid. p.70

[50] There is a footnote marked with a star '*', ibid. p. 70, which indicates that at this point Lewis had in mind Rescher's metalinguistic theory (Rescher 1964) rather than Goodman's theory.

[51] Cf. Lewis op. cit. section 2.3. on page 48.

permitting sphere throughout which $\phi \supset \psi$ holds, in which case $\phi \supset \psi$ is a premise, cotenable with ϕ at *i*, which together with ϕ implies ψ. Q.E.D.

If each of $\chi_1, ..., \chi_n$ is cotenable with ϕ, then so is their conjunction; so we can also say that $\phi \rightarrow \psi$ is true at *i* iff ϕ together with finitely many premises $\chi_1, ... \chi_n$, each cotenable with ϕ at *i* logically imply ψ.[52]

Note that Lewis allows the corresponding material conditional to be among the background propositions: if $\phi \rightarrow \psi$ is true and ϕ possible (entertainable), then $\phi \supset \psi$ can stand for χ. Goodman does not allow the corresponding material conditional to be among the background propositions – this is the change in his truth conditions that he made under the influence of Parry, and added to later editions of his original paper as Footnote 6[53]. He said there that "we must add the requirement that neither S nor S' follows by law from $\neg$A", that is, χ should not follow from $\neg\phi$. However, $\phi \supset \psi$ does follow from $\neg\phi$.

These are the facts. I will briefly repeat the main points about the relation between the two theories as Lewis described it, and then proceed with a counterexample to the notion of cotenability$_l$.

The auxiliary premises, or the background propositions, are inner necessary, provided the corresponding counterfactual is true. That is, if $\phi \rightarrow \psi$ is true, the χ's in virtue of which $\phi \rightarrow \psi$ is true are inner necessary. This means that there is a sphere throughout which all the χ's hold. That sphere is big enough to include some ϕ-worlds, provided ϕ is possible.

---

[52] Ibid. p. 57.

[53] I mentioned this footnote in Chapter 2 on Goodman. See pp. 15, 19-20 above.

Let us consider 4.1 again in the same context, i.e. the match m is dry, well made, and in presence of enough oxygen, and all such matches light when struck (χ). m is not struck ($\neg\phi$) and it doesn't light ($\neg\psi$). Why is 4.1 true? What we want from a philosophical theory is to provide *explanation.* Thus good semantics would not only tell us what the truth value of a conditional is, but it would also explain why it has that truth value.

According to the metalinguistic theory, 4.1 is true because the argument from ϕ and the background χ to ψ is valid. According to Lewis's theory, 4.1 is true because at the most similar worlds where I strike m, m lights. Why does it light there? Because, according to what Lewis says about the connection between his and the metalinguistic theory, at the most similar worlds where I strike the match, it is also dry, well made, etc. That is, the relation of similarity makes an ordering of worlds on which the χ's are inner necessary, and they hold throughout a sphere big enough to include some ϕ-worlds. Thus the closest ϕ-worlds are also χ-worlds, and they all must be ψ-worlds because ϕ and χ logically imply ψ.

Let us now consider some more propositions in the same context. Let ξ = I put m into water, and $\neg\chi_1$ = m is wet. Both ξ and $\neg\chi_1$ are false. The following conditional is also obviously true:

4.2 Had I put m into water, it would have been wet. ($\xi \rightarrow \neg\chi_1$)

Suppose further that ϕ and ξ are irrelevant to each other, in that the match does not have to be put in water to be struck, and vice versa. Trivially, ϕ-worlds and ξ-worlds are either equally similar to *i*, or one group of worlds is more similar than the other.

Intuitively, this has nothing to do with the truth of 4.1 and 4.2. In other words, 4.1 and 4.2 are true in the described context no matter whether $[\phi]<_i[\xi]$ or $[\xi]<_i[\phi]$.

Suppose $[\xi]<_i[\phi]$, i.e. my putting m into water is closer to actuality than my striking it. For example, I might be standing up to my knees in water, and holding dry, well made m in the presence of enough oxygen. The nearest matchbox is miles away, so there is nothing appropriate to strike the match on. It is less of a departure from actuality that I just put my hand into water than that something appropriate appear on which to strike the match. Since "m is dry" (χ_1) is one of the χ's in virtue of which 4.1 is true, according to Lewis χ_1 should be inner necessary and it should hold throughout a ϕ-permitting sphere. However, since 4.2 is also true, m's being put into water and it's being wet ($\xi\wedge\neg\chi_1$) is closer to actuality than it's being struck (ϕ). Thus χ_1 cannot hold *throughout* some ϕ-permitting sphere, which makes it not cotenable$_l$ with ϕ. (See figure 3.)

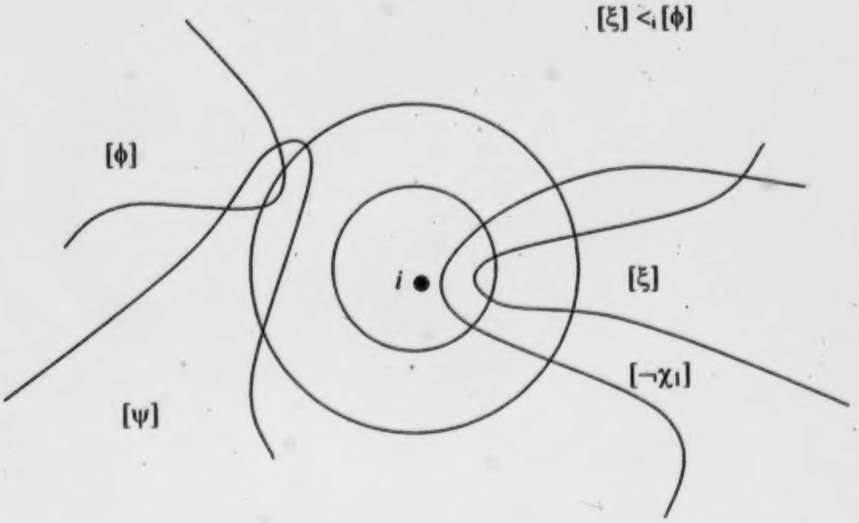

Fig. 3.

The general pattern for making similar counterexamples goes as follows. We need three truths of the form: $\phi \rightarrow \psi$, $\xi \rightarrow \neg\chi_1$, and $[\xi] <_i [\phi]$, where χ_1 is one of the auxiliary propositions that make $\phi \rightarrow \psi$ true, and where ξ and ϕ are irrelevant to each other.

What can we infer from this counterexample? The match's being dry (χ_1) turned out not to be cotenable with its being struck (ϕ). This makes Lewis's notion of cotenability inadequate, since we need χ_1 among the background propositions. It is not inadequate just because it is different from Goodman's notion, or because it is not in accordance with our natural language understanding of the word, but because it is useless and has no explanatory power. According to Definition 5, since χ_1 is not cotenable$_l$ with ϕ, it may not appear in the argument of the form

$$\frac{\phi, \chi_1, \ldots, \chi_n}{\psi}$$

What are the background χ's that do appear in the argument, then? They are not the propositions called $\chi_1, \ldots, \chi_4$ in my example. But then if $\phi \rightarrow \psi$ is true according to Lewis's theory, Definition 5 cannot play any role in an explanation of why this is so. And, as we said above, this is one of the main tasks of a theory of conditionals – to give an account of *why* conditionals have the truth values that they have. Having some arbitrary propositions among the χ's in the above argument says nothing about why 4.1 is true. For example,

$$\frac{\phi, \phi \supset \psi}{\psi}$$

is a valid argument and $\phi \supset \psi$ is cotenable$_l$ with ϕ, but this does not tell us much about 4.1. Modus ponens's being valid for '$\supset$' is not too informative. Neither is $\phi \supset \psi$ being cotenable$_l$ with ϕ, because it only says *that* $\phi \supset \psi$ holds throughout a ϕ-permitting sphere, without explaining *why* it is so.

So far, we can make the following conclusions:

(a) Lewis's notion of cotenability is inadequate, and

(b) Definition 5 is inadequate.

This raises a more serious worry. Definition 5 is equivalent to Definitions 1 – 4, so if something is wrong with one of them, the same thing is wrong with the others. What makes Definition 5 inadequate is Lewis's notion of cotenability. That notion is not primitive for Lewis. This indicates that the problem expressed in terms of cotenability$_l$ is also a problem for the primitive notion, similarity. In other words, we cannot just forget about Lewis's definition of cotenability and Definition 5, while keeping the rest of the theory as it is.

Let us try to see what the counterexample says about similarity. We saw that χ_1 does not hold throughout a ϕ-permitting sphere. Now the question is whether χ_1 holds at the closest ϕ-worlds, i.e., whether m is dry at the closest worlds where it is struck. Lewis's theory does not tell us if it is or it isn't. That is, according to everything we can know about Lewis's theory, we cannot answer. What in Lewis's theory ensures that χ holds at the closest ϕ-worlds is the inner necessity of χ. Now that we see that χ is not inner necessary, we don't know any more what is going on at the closest ϕ-worlds.

This means that Definitions 1 – 4 do not explain why the conditional 4.1 has the truth value that it has. The usual explanation Lewis's theory would offer is that 4.1 is true (if it is true) because m lights at the closest worlds where it is struck. When we consider my example we can see that this explains nothing, but rather can only confuse us, because we don't know whether m is dry at those worlds. Claiming that we *do* know that the match lights and that we do *not* know whether it is dry is not an explanation. And, similarly, if it were not the case that the closest ϕ-worlds were ψ-worlds, this would not provide an explanation of 4.1 being false.

Can we just assume that the χ's hold there? In the next chapter I will discuss Lewis's possible ways out of the 4.1 - 4.2 problem. For now we can say that it is not clear whether the assumption can be made. Some worlds where m is wet are closer to the central world than any world where m is struck. Hence there is no reason to suppose that $[\phi\wedge\chi_1] <_i [\phi\wedge\neg\chi_1]$ *must* be true (that the worlds where m is dry and struck must be more similar than the worlds where m is wet and struck). In any case, the assumption is not Lewis's. Adding the assumption is something we would need to work on to improve the theory. It is not trivial whether the assumption would help or not. So we can infer about Lewis's theory *as it is now* that

(c) Definitions 1 – 4 are inadequate, as well as Definition 5.

So far we see that the theory lacks the explanatory power it must have in order to be acceptable, but we do not know yet what makes the problem. Here is a suggestion. We saw that Lewis's theory works perfectly when we evaluate only one conditional. In the case of 4.1, we could adjust the similarity relation so that m's being struck (ϕ) was closer to actuality than its being wet ($\neg\chi_1$),. Thus we would *not* notice the lack of explanatory power I described above. The counterexample and the pattern for making similar examples show that there are many cases where for one counterfactual there is another that cannot be evaluated on the same scale of similarity. When we consider 4.1 by itself, everything is fine, but when we consider 4.1 together with 4.2, everything gets confused and we don't know any more how to understand the similarity relation. This suggests that

(d) similarity should not be 'overall' or 'absolute'; it should be made relative in some way, so that it can be adjusted when evaluating multiple conditionals.

(In what way should it be relative? This is not an easy question. I will discuss that later.)

It is worth noting that the metalinguistic theory is *equally* good (or bad, if you want) at explaining 4.1 and 4.2 separately and together. Lewis's theory is (at least) as good as the metalinguistic theory at evaluating and explaining 4.1 by itself, but it fails in considering 4.1 and 4.2 together. If my conclusions are right, than it obviously follows that

(e) there is no connection of the kind Lewis describes between his and the metalinguistic theory.

More will be said about Lewis's theory in the next chapter. I will try to make more general conclusions that do not depend on features peculiar to Lewis's theory, but pertain to a larger class of possible worlds semantics for counterfactuals. So I'll say more about Lewis under the title "Absolute similarity".

Let us now compare the usual interpretations (UI's) of Goodman once again. The set S of background propositions must contain the propositions relevant for the counterfactual whose truth value is evaluated. Nute's version of UI puts into S besides the relevant proposition all irrelevant ones that do not make problems, i.e. all propositions cotenable with the antecedent. That makes valid conditional excluded middle, because an antecedent and everything cotenable with it determine a single world. Lewis's version of UI says that the number of background propositions is finite. That way CEM is not valid, because a finite number of propositions cannot determine only one world (unless it contains infinite conjunctions, which is not intended to be the case). Why did Lewis say that the number must be finite? Did he think that this was correct interpretation of the metalinguistic theory? Or did he realise that allowing S to be infinite he won't be able to relate his semantics to the metalinguistic theory (because of CEM)? I don't know. He did not explain. He might have been led also by a purely formal reason. At one point he talks about conjunction of background propositions (see the proof of equivalence of Definition 5 and Definitions 1 – 4 cited above). According to the standard definition of a logical formula, the number of conjuncts in a conjunction must be finite. Otherwise, it is not a formula.

Whatever the answer might be, let us try to see what is to be included among those finite number of propositions. Apparently, these are not only the relevant propositions. There are worlds where m is struck, well made etc. ($\phi \wedge \chi$), but where the law of gravitation is different, the sun is the only star, and all presidents are smart. It is important that ψ holds at such worlds also. Otherwise, the argument from ϕ and χ to ψ would not be valid. However, it is not likely that Lewis would include such worlds among the closest ϕ-worlds. Their 'overall' similarity places them far away from the central world. The closest ϕ-worlds, as we saw, are those where $\phi \wedge \chi$ holds. If S contained only the relevant propositions, then all sorts of weird worlds would count as closest ϕ-worlds, as long as both ϕ and χ hold there. Thus S should contain, beside the relevant propositions, *the important* truths, such as the actual law of gravitation and the like. Important in what sense and for what? Well, important, if not for the counterfactual we evaluate, then for the overall similarity. If you understand the notion of overall similarity, then you know what is to be included in S.

(Please remember that Lewis's UI is not Lewis's theory, nor probably even his exact interpretation of Goodman or some other version of the metalinguistic theory. The purpose of the last few paragraphs was to try to relate Lewis's UI to the similarity theories.)

5. Absolute similarity

In chapter 7 I will propose a new theory of counterfactuals. I am therefore obliged to compare it to other theories. Of course, this is practically possible to be done only by comparing it to types of theories, rather than particular theories, whose number is huge. My job is made considerably easier by Nute's very convenient classification of possible worlds semantics for counterfactuals[54]. He also axiomatized various theories in a way that makes them easy to compare. I benefited from that throughout this thesis.

Possible worlds semantics use as a primitive notion a similarity relation or a selection function (or can be reformulated that way). The purpose of that primitive notion is to somehow distinguish between important and unimportant worlds, where 'important' in great majority of cases means 'more similar'. Various theories give a different answer to the question 'How similar are the important worlds?'. That is, the selection function can be more or less inclusive. It can include only the most similar world(s), or those that are 'similar enough', or those that are similar only in some respects, while they can be very different in all others. That is the basis for Nute's classification. He distinguishes between minimal, small, and maximal change theories.

Besides Stalnaker's and Lewis's, other minimal change theories that Nute mentions were proposed by Pollock[55], Kratzer[56], Blue[57] and Veltman[58]. The notion of minimal change is related to what Lewis calls 'limit assumption', an assumption that for any possible antecedent there are always antecedent-worlds closer to actuality than any other antecedent worlds. The selection function in minimal change theories selects those closest antecedent-worlds. We saw (end of chapter 3 above) that in Lewis's semantics the limit assumption does not hold: there can be antecedent-worlds that endlessly approach some degree of similarity without ever reaching it, so that no closest antecedent worlds exist. In other words, there is no smallest antecedent-permitting sphere. That is why Lewis uses the notion of comparative similarity instead of selection function: instead of checking whether the consequent holds at the closest antecedent-worlds, he compares the antecedent-worlds where the consequent holds and the antecedent-worlds where the negation of the consequent holds. Still similarity-based semantics and selection function-based semantics, as Nute says, give the same truth conditions for counterfactuals in the cases where the closest antecedent-worlds do exist. That is why Nute classifies Lewis's theory as a minimal change theory, to distinguish it from theories that lack that feature[59]. It is not clear to me whether within Lewis's semantics we can ever talk about the most similar worlds at which some false antecedent holds. But there are more reasons to put Lewis's theory together with other minimal change theories. Informally speaking, Lewis's theory tends to come as close as possible to the minimal change. Smaller change is always more relevant than a small change, and less relevant than still smaller change. In other words, what is going on in smaller antecedent-permitting spheres is generally more important than what is going on in small antecedent-permitting spheres. This is reflected in the fact that adding the limit assumption to Lewis's semantics has no influence on the formal system: the set of theorems remains the same with or without the limit assumption[60]. Thus the semantics for Lewis's favourite system for counterfactuals that he calls VC[61] can be expressed in terms of a selection function instead of the similarity relation. The selection function f for VC, which takes as arguments a proposition (antecedent) and a world, satisfies the following restrictions[62]:

---

[54] Nute and Cross 2002.
[55] Pollock 1976, revised in Pollock 1981.
[56] Kratzer 1979 and 1981.
[57] Blue 1981.
[58] Veltman 1976.

---

[59] Cf. Nute and Cross 2002 p. 15.

[60] Cf. Lewis 1973. p. 121.

CS1 if $j \in f(\phi, i)$ then $j \in [\phi]$

CS2 if $i \in [\phi]$ then $f(\phi, i) = \{i\}$

CS3 if $f(\phi, i)$ is empty then $f(\psi, i) \cap [\phi]$ is also empty

CS4 if $f(\phi, i) \subseteq [\psi]$ and $f(\psi, i) \subseteq [\phi]$, then $f(\phi, i) = f(\psi, i)$

CS5 if $f(\phi, i) \cap [\psi] \neq \varnothing$, then $f(\phi \wedge \psi, i) \subseteq f(\phi, i)$

CS6 $i \in [\phi \rightarrow \psi]$ iff $f(\phi, i) \subseteq [\psi]$

The second condition CS2 says that if an antecedent is true at i, then i is the ϕ-world closest to i, and all other ϕ-worlds are less similar to i. The function f in that case selects only i. The syntactic counterpart of CS2, which holds in all minimal change theories, and which many philosophers find objectionable[63], is

CS $(\phi \wedge \psi) \supset (\phi \rightarrow \psi)$

Replacing conditional excluded middle (CEM) in Stalnaker's system C2 (see pages 24-5 above) with CS would give us Lewis's system VC.

Nute[64] proposed a *small change* theory based on a selection function whose intended interpretation is to pick up not only the closest worlds, but also those that are 'similar enough'. That way the problem with the limit assumption is avoided, because $f(\phi, i)$ for possible ϕ's is never empty, even if there are no closest ϕ-worlds. The other difference from Lewis's system is that such a function cannot select only one world for a true antecedent, because besides the actual world, there are many other worlds that are 'close enough'. Thus in Nute's semantics CS2 is replaced with

---

[61] Cf. ibid. chapter 6.

[62] CS1-6 are from Nute and Cross 2002, p. 15.

[63] The first objection to CS that I know of is in Bennett 1974.

[64] Nute 1975a, 1975b, 1980

CS2' if $i \in [\phi]$ then $i \in f(\phi, i)$

and hence CS is not a theorem.

Other motives to adopt small change theories are related to some problems with the limit assumption. They will not be discussed here. I will eventually reject the notion of similarity as misleading, and therefore do not want to be involved in a discussion on the limit assumption more than it is needed to explain the classification proposed above. Any discussion on that topic is, I believe, more about similarity than about counterfactuals. Other small change theories that Nute mentions were proposed by Åqvist[65] and Warmbrōd[66].

Gabbay[67] argued that conditional logic based on a selection function of the form $f(\phi, i)$ cannot match our ordinary language practice. The function should have another argument besides ϕ and and i, and that is the consequent. That way Gabbay's function selects different worlds for counterfactuals that have the same antecedents but different consequents, unlike any possible worlds theory we mentioned so far. The intuition behind this technical solution is that 'overall' similarity is not what matters. Only some limited respects of similarity are important – those relevant for the relation between the antecedent and the consequent. In all other respects selected worlds can differ from i to any degree whatsoever. For that reason Nute calls Gabbay's theory a *maximal change* theory. Other theories of this kind were proposed by Nute and Fetzer[68].

---

[65] Åqvist 1973.
[66] Warmbrōd 1981. Warmbrōd's theory will be analyzed in details later (chapters 6 and 7). It is not clear that the theory is to be classified as small change. I would rather classify it as minimal change.
[67] Gabbay 1972. and 1976.

To Nute's classification I add another one. I will distinguish between theories based on an *absolute* similarity or selection function and those based on a *relative* similarity or selection function. Bennett explains the notion of absolute similarity:

> In Lewis's 'Ptolemaic astronomy' of possible worlds, all the worlds lie in concentric shells around α with each one's degree of closeness to α being represented by its distance from the centre. These closeness relations are fixed independently of what conditional one is evaluating. For Lewis, as for Stalnaker, standards of closeness can vary with context; but they do not vary according to what the antecedent is of the conditional in question.[69]

... nor do they depend in other ways on the conditional or conditionals that we are evaluating. In different contexts we may choose different ordering of worlds; but once we make our choice, all counterfactuals are to be evaluated according to that ordering. In the next chapter I will discuss theories where the selection function is relative either to the antecedent, or both the antecedent and the consequent, or context. Context relativity is different from the context-dependence of absolute similarity. The former uses a selection function that applies only to conditionals that belong to the same context or piece of discourse; the latter, as we said, applies to all conditionals.

In this chapter I will present some more arguments against Lewis's theory, but now I want to draw more general conclusions with the aid of the two classifications proposed above. The arguments that follow do not depend on any specific feature of Lewis's theory except these three. Lewis's theory is

---

[68] Fetzer and Nute 1979, 1980. Nute 1981.

[69] Bennett 2003 p. 298f.

- minimal change
- based on absolute similarity
- based on an assumption that any two worlds can be compared according to their similarity to a third world.

So the conclusions drawn in this chapter pertain to any such theory. They also pertain to any *small* change theory that has the last two features listed above. The conclusions do not depend on whether similarity or selection function is used. Therefore, in this chapter I argue against any *total ordering minimal or small change theory based on an absolute similarity or selection function.*

My critique of this kind of theories, and some others in the next chapter, involves a comparison of these theories to the metalinguistic theory. Before applying that method, let me try to justify it.

The superiority of possible worlds semantics is obvious, because it makes it much easier to build logical systems than the metalinguistic theory does. Why bother, then, with the metalinguistic theory? Why worry about these old stories now that more powerful logical tools are available? Because the metalinguistic theory has more than merely historical importance.

Let us consider the match example again. Suppose I claim A→C, i.e.

Had I struck m, it would have lit.

The context is the same as before, i.e. m is dry, well made etc (that is, the propositions B_1 – B_4 from the Chapter 2 hold). Now you ask for an *explanation.* Why is A→C true? A philosophical theory (*qua* philosophical) must offer an answer to the question *why* a conditional is true or false, not only to claim that it is so. What would you expect as an answer? What answer or what kind of an answer would be satisfactory? Well, listing the background propositions $B_1 - B_4$ is a reasonably good explanation. Usually we are satisfied with such an answer. The match would light if struck because it is dry, well made, etc. and all such matches light when struck. This is the explanation that the metalinguistic theory offers.

What explanation does possible worlds semantics offer? Possible worlds semantics for counterfactuals, of Lewis's or any other kind, tell us that in evaluating the truth value of a conditional we should pay attention to some worlds and ignore others. Some worlds are important and the rest are not. $A \rightarrow C$ is true because C holds at the important worlds. There are different opinions on what counts as important, and because of that different similarity relations and different selection functions have been proposed. Their role is to somehow separate important from unimportant worlds. The similarity relation has a task of ordering the worlds in such a way that the important worlds come closer to the central world than the unimportant worlds. The task of the selection function is to select (only) the important worlds.

Now we can state explicitly what is the importance of the metalinguistic theory. To make conditional logic we use possible worlds semantics because they give us more powerful logical tools than the metalinguistic theory does. Nevertheless, we use (or should use, as I believe) the notion of background propositions to test the adequacy of the notions of similarity or selection function. In order for possible worlds semantics to have explanatory power and to make sense, the following interpretation of the notion of 'important' worlds must be *possible*:

5.0 *The relevant background propositions are true at the important worlds.*

An explanation of why A→C is true offered by a possible worlds theory is that the conditional is true because C holds at the important worlds. Whatever we may otherwise assume by 'important', a necessary part of the meaning of that notion is that the relevant background propositions hold at all the important worlds. An interpretation of the selection function or similarity that is in accordance with 5.0 must be *allowed.*

For example, minimal change theories say that A→C is true because C holds at the most similar A-worlds. Part of the meaning of 'the most similar worlds' is that the relevant facts hold at those worlds. In the match example, we are interested in the worlds where the match is dry, well made, etc, i.e. where $B_1 - B_4$ hold, and where the antecedent is true as well. Obviously, we are not interested in the worlds where the match is wet. Thus a selection function that works properly would not select any worlds where the match is wet, or where any of other B's does not hold.

What happens if the connection between the metalinguistic and the possible worlds theories as required by 5.0 cannot be established? Without this connection similarity is an abstract notion without meaning, or (even worse) a notion that is misleading because it merely appears to be meaningful. If this connection doesn't hold, our semantics would have no explanatory power. Similarity or selection function becomes something like Hilbert's ideal elements, and gives no meaning to counterfactuals. It is no more a philosophical theory, but merely an exercise in model theory.

In the previous chapter we saw that Lewis's notion of cotenability does not pick up the propositions we would expect, namely the relevant background propositions. In this chapter I will try to show why it doesn't pick them. Because it can't (as well as any other notion of cotenability defined in terms of absolute similarity couldn't). And it can't because 5.0 does not hold in Lewis's theory. If we suppose that it does hold, we fall into contradiction.

While reading Lewis I got an impression that he wanted to describe the relation between his and the metalinguistic theory as being in accordance with 5.0. Let me explain what I will call Lewis's *informal* understanding of the notion of similarity, which *is* in accordance with 5.0, and which I think is implicit in his description of the relation between his and the metalinguistic theory.

Let us see what are the important worlds according to Lewis, first according to his formal semantics, and then according to his informal understanding of similarity. When we evaluate a conditional $\phi \rightarrow \psi$ the semantics tells us to look at the first ϕ-worlds we meet when we move away from the central world *i* according to the assumed ordering of worlds[70]. Lewis's *informal* explanation of similarity tells us what those worlds are. As we saw from his explanation of the relation between his and the metalinguistic theory, if $\phi \rightarrow \psi$ is true, then the following are also true:

---

[70] "The first worlds we meet" is Bennett's trick to avoid saying "the closest worlds", because there might be no closest worlds in Lewis's semantics. No harm will be done if we assume that the trick works and do not think what version of Zeno's paradoxes may come out of it. If we want to be more precise, the important worlds for Lewis might be defined as either (i) $A \wedge C$ worlds that are closer than any $A \wedge \neg C$ worlds, if $A \wedge C \prec_i A \wedge \neg C$, or (ii) $A \wedge \neg C$ worlds that are closer than any $A \wedge C$ worlds, if $A \wedge \neg C \prec_i A \wedge C$, or (iii) A-worlds that belong to some sphere where for each $A \wedge C$-world there is an equally similar $A \wedge \neg C$-world and vice versa, if $A \wedge C$-worlds and $A \wedge \neg C$-worlds are tied, or (iv) the central world only, if A holds at it. (i)-(iii) assume that A does not hold at *i*. I will use Bennett's trick to avoid this clumsy sentence.

(3) $[\chi] <_i [\neg\chi]$

(4) $[\chi \wedge \phi] <_i [\neg\chi]$

(5) $[\chi \wedge \phi] <_i [\neg\chi \wedge \phi]$

where χ is a conjunction of all the relevant background propositions $\chi_1, \dots \chi_n$ in virtue of which $\phi \rightarrow \psi$ is true. These formulae express what I call Lewis's informal explanation of the similarity relation. (3) and (4) are explicitly stated in Lewis's words we cited above, namely that χ is inner necessary and the scope of its necessity is big enough to include some ϕ-worlds. (5) is entailed by (4). It says that the antecedent-worlds where the background propositions hold are closer than the antecedent-worlds where the background propositions do not hold.

Formulae (3) – (5) give us an important part of the *meaning* of the word 'similarity' in the context of Lewis's semantics. They tell us what the worlds that are important are. In other words, they tell us that the first ϕ-worlds we meet when moving away from *i* are those where both ϕ and χ hold. This idea looks very attractive and makes perfect sense. In evaluating 4.1 (Had I struck the match m, it would have lit: ($\phi \rightarrow \psi$)), for example, the important worlds are those where the match is dry (χ_1), well made (χ_2) etc. The worlds where it is not dry are obviously not important. The role of the similarity relation is to separate important from unimportant worlds. That relation is supposed to order the worlds in such a way that the important worlds come closer to *i* than the unimportant worlds. That is why the worlds where the match is dry, well made etc. are more similar than those where this is not the case.

However, Lewis's formal system is not in accordance with what I called his informal understanding of similarity. We saw in the counterexample from the last chapter (page 51 above) that (4) (and hence (5) as well) is false while 4.1 is true. There are more such examples when $\phi \rightarrow \psi$, $\xi \rightarrow \neg\chi_1$, and $[\xi] <_i [\phi]$ are all true. These counterexamples are possible because cotenability$_1$ does not pick up the relevant background propositions. Trying to find another definition of cotenability in terms of absolute similarity would not help because, as I said, the semantics does not obey 5.0. Adding 5.0 leads to a contradiction.

Fact Each truth (at the actual world *i*), say χ, can be a background proposition for some counterfactual $\phi \rightarrow \psi$, no matter how distant from *i* the closest ϕ-worlds are.

In other words, each actual truth can be (part of) a reason, or a cause, or an explanation of why some ψ would hold had some ϕ been the case (no matter how unlikely it is for ϕ to be the case). Let us consider some examples to make this claim more clear. Pick a truth, say, "You are now reading my thesis", and call it χ. *Fact* says that there is a conditional that is true because χ is true. Here is one: "If X were to look through your window, he would see you reading something". *Fact* also says that that the antecedent-worlds can be arbitrarily distant from the central world. Replacing X with different individuals we can get pretty far away from actuality. For example, X can be your neighbour, a person now in Chukotka[71], a person now in a Tur[illegible]sh prison, Aristotle, a two-headed Aristotle, etc. In general, we can pick any truth and, with a bit of imagination, find a counterfactual that is true in virtue of the truth we pick.

---

[71] I bet you are not in Chukotka now.

If this is so, the problem is that every actual truth has to be inner necessary, with an arbitrarily large scope of necessity (because, as we saw, Lewis said that every background proposition is inner necessary, with the scope of its necessity big enough to include some antecedent-worlds). But this is contradictory. In that case, whichever direction we move from the central point representing the actual world, we meet worlds where exactly the same propositions are true as in the central world. What then would be the difference between those worlds and the central world? They turn out to be the same, but we assumed that they were different worlds.

From this contradiction we can infer that absolute similarity cannot be adjusted to all counterfactuals. If we order the worlds in the way that gives the right truth values for some conditionals, the same order gives wrong truth values to some other conditionals. We can give particular examples, and a general pattern that generates such examples, that lead to the same conclusion.

Let us consider again the situation with the match m from the counterexample presented in the last chapter. Again let χ be true and ϕ, ψ and ξ (I put m into water) false. For the same reasons as before the conditional 4.1

4.1 Had I struck the match m, it would have lit ($\phi\rightarrow\psi$)

is true. Suppose further that for whatever reason I examine a pile of matches in order to strike *all and only* wet ones and see if they would light (say, I got a grant from my university to do that research). The match m is at my disposal, so it is true that

5.1 Had m been wet, I would have struck it ($\neg\chi_1\rightarrow\phi$)

Suppose also that it is more possible that m becomes wet than to be struck while still dry. For example, there is some water around, but there is nobody around to strike the match, because I strike only wet matches. Therefore $[\neg\chi_I\wedge\phi]<_i[\chi_I\wedge\phi]$ is true, which contradicts (5). The situation looks like:

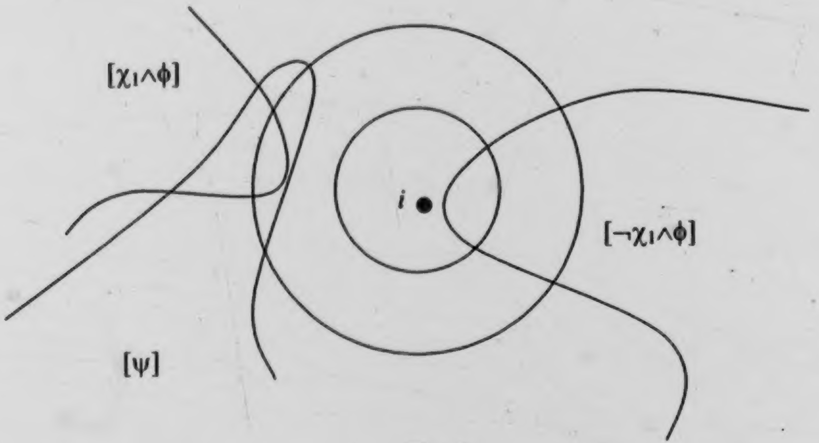

Fig. 3.

The closest ϕ-worlds are $\neg\psi$-worlds (the match is wet there, so it doesn't light). This makes 4.1 false in Lewis's semantics. However, this is intuitively wrong. My having that grant has nothing to do with 4.1, so there is nothing in the example that makes 4.1 false. One might try to explain 4.1's being false by saying that I would only have struck m if it were wet. However, this is backtracking reasoning, which Lewis – rightly, I believe – does not allow for counterfactuals. When evaluating 4.1 we are interested in ϕ-worlds at which χ holds. $\phi\wedge\neg\chi$-worlds should not be relevant.

The general pattern for making this kind of counterexamples goes as follows. We need three truths of the form $\phi\rightarrow\psi$, $\neg\chi_I\rightarrow\phi$, and $[\neg\chi_I]<_i[\chi_I\wedge\phi]$, where χ_I is one of the background propositions that make $\phi\rightarrow\psi$ true. According to Lewis's semantics a set containing these three propositions has to be inconsistent. However, this is counterintuitive. The set should be satisfiable.

Near the end of the last chapter, in my point (d) (page 54 above) I suggested that similarity should be relative, not absolute. There are different kinds of relativity proposed in the literature. Similarity can be antecedent-relative, or antecedent-and-consequent relative, or relative to context. In Lewis's terms this would mean that for each antecedent (or antecedent-and-consequent, or context) we need a different ordering of worlds. That way conditionals with different antecedents (or antecedent-and-consequent, or context) could not be evaluated in the same system of spheres. Thus 4.1 and 5.1 could not be evaluated in the same system of spheres. This is a possible solution. Making similarity relative in a way that prevents 4.1 and 5.1 to be evaluated according to the same ordering enables us to adjust the similarity relation to both conditionals and have the right truth values for them.

6. Relative similarity

6.1 Antecedent-relativity

The conclusion from the previous chapter suggests that instead of an absolute similarity or absolute selection function we should use a relative notion of similarity or selection function. The examples above describe contexts where a pair of intuitively obviously true conditionals cannot both be true in a single model within Lewis-type semantics. These conditionals do not 'go together'. They require different criteria of similarity and different ordering of worlds. In Lewis's terms, they require different systems of spheres. We cannot make an ordering of worlds that would be good for all counterfactuals at the same time. We cannot decide which features of similarity to accept and which to ignore, because our intuitions about counterfactuals require too many incompatible features of similarity to hold at the same time. Thus 4.1 and 5.1 from the previous chapter require different orderings, and the general pattern for such examples tells us that there are many such conditionals that can't go together. This also holds for the pair 4.1 and 4.2 from the counterexample to cotenability$_l$, although the pair 4.1 – 5.1 leads to that conclusion more obviously.

An intuition behind these claims might be this. If we are considering what would happen if I stuck the match, we want to know what happens in the antecedent-worlds where the match is dry, well made etc., and we call these worlds more similar than those where the match is wet or not well made. These worlds where the match is dry, well made etc. still differ from our world in some respects. In what respects? Well, the first answer that comes to mind is that they differ in some respects irrelevant for the case with the match m. But how are we going to determine what is irrelevant absolutely? In these worlds, for example, I can be taller than I actually am. This is irrelevant for evaluating the truth value of the conditional 'Had I struck the match it would have lit'. However, it is not irrelevant for evaluating for the conditional 'Had I decided to make a career as a basketball player...'. In whatever respect those worlds differ from ours, it is going to affect some (very many) other counterfactuals and our criteria of similarity is going to give us the wrong truth value for them. That is why they should not be considered in the same system of spheres.

However, let us recall Goodman's problem of choice between 2.2 and 2.3 (were the match struck would it light or be wet?). At the beginning of the Chapter 3 we introduced the basic idea of minimal change by showing the way it solves the 2.2 – 2.3 problem, without getting into trouble as Goodman did. The solution looked very elegant and attractive: we introduce only the necessary changes that allow the match to be struck, and we see that after those changes the match would not be wet, because its becoming wet is not necessary for its being struck. And it intuitively looked very plausible to generalize from such cases and define the truth conditions for counterfactuals thus:

3.0 A→C is true iff C will be true after only the changes necessary to make A true have been made.

After that we saw that the minimal change theories of Lewis's kind do not work. What is wrong then with such an appealing and elegant solution as 3.0?

It seems that whatever was so attractive about 3.0 was not properly captured by the theories we discussed above. To evaluate a conditional we consider only the necessary changes that make the antecedent true. How does the idea of absolute similarity follow from that? It does not. The idea of minimal change (which is usually assumed to be expressed by 3.0) does not necessarily require the notion of absolute similarity. In evaluating the conditional 'Had I struck the match...' we worry about the closest worlds where the match is struck. In evaluating the conditional 'If kangaroos had no tails...' we worry about the closest worlds where the kangaroos are without tails. And so on. What worlds are important seems to depend on the antecedent. Different things are important in the case where we consider what would happen if I struck the match and in the case when we consider what would happen if kangaroos had no tails. In the first case the kangaroos have no importance whatsoever, in the second the match is irrelevant. This suggests that similarity or selection function should not be relative only to the present states of affairs (i.e. the actual world and the context of utterance), as it is in the absolute similarity theories, but should be relative to the antecedent as well.

In Lewis's terms, this line of reasoning leads us to the conclusion that we need a different system of spheres for each antecedent, and that conditionals can be considered in the same system only if they have the same antecedent. Absolute similarity, in order to consider all counterfactuals in the same system and to be adjusted to all of them, must have incompatible properties. Antecedent-relative similarity is supposed to avoid that problem. In particular, semantics based on such a notion avoids all the problems with my examples above (the 4.1 – 4.2 and the 4.1 – 5.1 example), because conditionals in the examples that do not 'go together' have different antecedents, and therefore cannot be evaluated in the same system of spheres.

Antecedent-relative semantics are weaker than those based on an absolute similarity or selection function. An antecedent-relative semantics recently proposed by Johan Mårtensson[72] makes CSO and CV invalid:

CSO $((\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)) \supset ((\phi \rightarrow \chi) \equiv (\psi \rightarrow \chi))$

CV $((\phi \rightarrow \psi) \wedge \neg(\phi \rightarrow \neg\chi)) \supset ((\phi \wedge \chi) \rightarrow \psi)$

which hold in both Stalnaker's C2 and Lewis's VC (as was mentioned in Chapters 3 and 5). The semantic counterparts of CSO and CV in the selection-function version of VC are

CS4 if $f(\phi, i) \subseteq [\psi]$ and $f(\psi, i) \subseteq [\phi]$, then $f(\phi, i) = f(\psi, i)$

CS5 if $f(\phi, i) \cap [\psi] \neq \varnothing$, then $f(\phi \wedge \psi, i) \subseteq f(\phi, i)$

The reason for CSO and CS4 failing in the antecedent-relative semantics is this. The closest ϕ-worlds and the closest ψ-worlds can be different even though ψ holds at the closest ϕ-worlds and ϕ holds at the closest ψ-worlds, because we need one ordering of worlds for ϕ and another for ψ. Thus $(\phi \rightarrow \psi)$, $(\psi \rightarrow \phi)$ and $(\phi \rightarrow \chi)$ can be true and $(\psi \rightarrow \chi)$ false, because the first and the third conditional are evaluated according to one ordering, and the second and the fourth according to another ordering. CSO is especially important formula, because it expresses an essential property of absolute similarity. It is therefore also of essential importance for antecedent relativity that CSO does not hold.

The intuition behind CV and CS5 is this. If χ is cotenable (in Goodman's sense) with ϕ, then χ holds at least at some of the closest ϕ-worlds (i.e. $\phi \Diamond\!\!\rightarrow \chi$). Therefore the set of the closest ϕ-worlds includes as a subset the set of the closest $\phi \wedge \chi$-worlds. Thus whatever is counterfactually implied by ϕ is also implied by $\phi \wedge \chi$. In an antecedent-relative semantics this *might* not work, because $\phi \wedge \chi$ *might* require a different ordering of worlds than ϕ does. Antecedent relativity may have different versions. Some of them make CV valid, some do not.

---

[72] Mårtensson 2000.

My primary interest in this section is to test the intuition behind 3.0 in its antecedent-relativity version. The method I will apply is the same I used in the previous chapter – I want to see if antecedent relativity allows 5.0 (which says that the relevant background propositions must hold at the important worlds; see p. 64 above).

Fact 2 Each truth (say, B) that would not be false if A were true can be among the relevant background propositions for some true counterfactual A→C.

Trivial evidence is the case when $C = A \wedge B$.

Fact 2 follows from 3.0, 5.0, and the notion of minimal (and also small) change. If B would not be false had A been true, that is, if B is cotenable in Goodman's sense with A, and if B is true, then ¬B is not among the necessary changes required for A to be true. The A-worlds where B holds are therefore more similar to our world than $A \wedge \neg B$-worlds. This conclusion, but it also follows from the following:

From 5.0 and Fact 2 it follows that each truth that would not be false if A were true holds at the important worlds, i.e. at the closest A-worlds. Theorem 2 from Chapter 3 (pages 29-30 above) says that any proposition together with all the truths cotenable with it determines a single world. Therefore the set of closest A-worlds contains only one element. Hence this type of the minimal change antecedent-relative semantics validates CEM.

What can we do with this result? We can accept CEM, and then we are obliged to explain away the intuitions against that formula. Or we can reject 3.0, and then we are obliged to explain away why it looks so attractive; besides that, 3.0 seems to require minimal or small change, so rejecting 3.0 means that we have to reject the idea of minimal and small change, or somehow reformulate it in a way that does not require 3.0.

I think that we should reject 3.0. The explanation of what is going on, I believe, is that the antecedent relativity is not relative enough. We switched from absolute to antecedent-relative similarity because absolute similarity must have incompatible properties for it to be adjusted to all counterfactuals. Now we have a parallel problem that the antecedent-relative similarity must have incompatible properties to be adjusted to all counterfactuals with the same antecedent. An intuition behind these claims might be this. In evaluating A→C, e. g.

> Had I struck m, it would have lit.

we do not worry much about some truths that have nothing to do with the relation between striking and lighting, or striking and lighting of that particular match m in that particular context. For example, we do not worry about the fact that I can hear, i.e. that I am not deaf. However, we do have to worry about that in evaluating the conditional with the same antecedent

> Had I stuck m, I would have heard the characteristic sound of a match being struck.

Thus an actual truth (that I can hear) was irrelevant when we evaluated one conditional, but it become important when we evaluated another conditional with the same antecedent and a different consequent. In evaluating the first conditional the only actual truths that were really important were $B_1 - B_4$ (that m is dry, well made etc.), so we want these truths to hold at all the closest antecedent-worlds. Whether other actual truths hold at these worlds is less important. They may hold at some and not hold at other closest antecedent-worlds. However, this understanding of similarity, conjoined with the idea of antecedent-relativity, turned out to be bad, because it overlooked that an actual truth (that I can hear), even though it is not among the relevant background propositions $B_1 - B_4$ for the given counterfactual, should nevertheless hold at all the closest antecedent-worlds, the same as $B_1 - B_4$ do. This is because this truth is a relevant background proposition for another true conditional with the same antecedent. To put this in Lewis's terms, $B_1 - B_4$ are not the only propositions that should be inner necessary. The truth that I can hear should be inner necessary as well.

Thus we know that $B_1 - B_4$ are inner necessary, and we found a fifth proposition that should join them and hold at the closest antecedent-worlds. It is not hard to find a sixth such proposition, and seventh and eighth. And many more. So where does this stop? Is there any actually true proposition cotenable with A that should not be inner necessary? In other words, is there any such truth that cannot possibly be a background proposition for any counterfactual with the given antecedent A? It seems that there isn't any, and that is the point of Fact 2. Thus the similarity relation or selection function, in order to be adjusted to all counterfactuals with a given antecedent, can pick up only one world. Absolute similarity must have incompatible properties to be adjusted to all counterfactuals, i.e. it is an inconsistent notion. Antecedent-relative similarity is not an inconsistent notion, but it contradicts our intention to have more than one antecedent-world. It is because of the collision with that intention that I say that antecedent-relativity is not ‘relative enough’. Absolute similarity presupposes that all counterfactuals can be evaluated according to one similarity measure, i.e. in the same model or same system of spheres. That was wrong, as we saw, because some counterfactuals do not ‘go together’. Now we have a parallel problem: not all counterfactuals with a given antecedent ‘go together’.

6.2 Gabbay’s antecedent-and-consequent relativity

Long ago Gabbay proposed another kind of relativity generalizing from examples like:

> If I were the Pope, I would have allowed the use of the pill in India
>
> If I were the Pope, I would have dressed more humbly
>
> Clearly, in the first statement, we must assume that [after the changes that make the antecedent true] India remains overpopulated and poor in resourses, while in the second example nothing of the sort is required.[73]

These examples suggest that in determining what the ‘important’ worlds are, we should take into account not only the actual world, or only the actual world and the antecedent, but the consequent as well. Gabbay’s theory is based on an antecedent-and-consequent relative selection function. This kind of relativity avoids CEM, because his selection function does not have to select a set of worlds at which an antecedent and all truths cotenable with it hold (and which, as we saw, can have only one world as an element), but it selects a set of worlds relevant only to a given conditional.

Gabbay proposed a conditional logic based on a selection function that takes three things as arguments: a world, an antecedent and a consequent[74]. Here I will present a version of Gabbay's theory from Nute and Cross 2002[75]. A model for Gabbay's theory is a triple $< I, g, [] >$ where I is a set of possible worlds, [] is a function which assigns to each proposition ϕ a subset $[\phi]$ of I (all the worlds from I where ϕ is true), and g is a function which assigns to sentences ϕ and ψ and world $i \in I$ a subset $g(\phi, \psi, i)$ of I. A conditional $\phi \to \psi$ is true at i iff $g(\phi, \psi, i) \subseteq [\phi \supset \psi]$. Three restrictions are imposed on the function g:

---

[73] Gabbay 1976. p. 188.

[74] Gabbay 1972. That theory is included in Gabbay 1976, which I use for citations.

G1 $\quad i \in g(\phi, \psi, i)$

G2 $\quad$ if $[\phi] = [\psi]$ and $[\chi] = [\theta]$ then $g(\phi, \chi, i) = g(\psi, \theta, i)$

G3 $\quad g(\phi, \psi, i) = g(\phi, \neg\psi, i) = g(\neg\phi, \psi, i)$

Conditional logic G determined by this semantics is closed under three rules:

RCEC $$\frac{\phi \equiv \psi}{(\chi \to \phi) \equiv (\chi \to \psi)}$$

RCEA $$\frac{\phi \equiv \psi}{(\phi \to \chi) \equiv (\psi \to \chi)}$$

RCE $$\frac{\phi \supset \psi}{\phi \to \psi}$$

The intuitive sense of the function g is very different from any selection function we have considered so far. $g(\phi, \psi, i)$ does not pick up worlds similar to i, but only those at which certain conditions relevant for $\phi \to \psi$ hold. In the match m example, in evaluating $A \to C$, i.e. 'Had I struck m, it would have lit', g selects all the worlds (accessible from our world) where $B_1 - B_4$ hold (i.e. where m is dry, well made etc.). Some of those worlds must be very weird and unlike our world. There can be a world where m is dry, well made, etc. but the law of gravitation is different and sun is the only star. For that reason Nute calls this type of semantics a *maximal* change theory.

---

[75] Nute and Cross pp. 25-6.

We saw that CSO and CV did not hold in one version of antecedent-relative semantics (pp.74-5 above). They do not hold in G either, for parallel reasons. CSO is an important formula because it holds in most theories of counterfactuals, and its semantic counterpart CS4 expresses some important features of absolute similarity, features that are not characteristic of relative similarity. It also appears to be very plausible. Do then relative semantics score a negative point for making CSO invalid? I turn to that question now.

6.3 CSO

Mårtensson attacked CSO ($((A\rightarrow B)\wedge(B\rightarrow A)\supset((A\rightarrow C)\equiv(B\rightarrow C))$) by presenting a counterexample to a formula that is a consequence of CSO, and that he calls RCV[76]

RCV $((A\rightarrow C)\wedge(A\rightarrow B))\supset(A\wedge B\rightarrow C)$

I will try to attack CSO directly. Again,

4.1 Had I struck m, it would have lit. $(A\rightarrow C)$

and, as before, $B_1 - B_4$ are true, and so is 4.1. A and C are false. What new happened in the meantime is that I got another grant, this time to strike all and only matches that light. The rules are strict, so I will lose my job if I strike a match that is not burning. Thus

6.1 Had the match m lit, I would have struck it $(C\rightarrow A)$, and

6.2 Had I struck m, I would have lost my job $(A\rightarrow D)$

are true, but, contrary to what CSO predicts, the following is false:

6.3 Had the match m lit, I would have lost my job $(C\rightarrow D)$

---

[76] Mårtensson 2000 p. 53.

Here is another counterexample (since you might have had enough of the match m perversions). Pressing the button on my lamp turns the light on or off. When the light goes on, the lamp plays the anthem for a minute. This does not happen when the light goes off. It's around midnight, I want to sleep, so the light would bother me. The light is off.

6.4 If I pressed the button, the light would be on. (A→C)

6.5 If the light were on, I would press the button. (C→A)

6.6 If I pressed the button, I would hear the anthem. (A→D)

But not

6.7 If the light were on, I would hear the anthem. (C→D)

What you might find wrong about these counterexamples is that they do not treat propositions as fragile entities. If we understand A, C, and D as fragile, that is, pertaining strictly to one moment of time, than the counterexamples actually contain more than three atomic propositions. There is a time gap between the antecedent-event and the consequent-event in all four counterfactuals involved. Then if A→C and C→A are true and the two C's are at the same time t, then the two A's are in fact different propositions, one referring to an event before t, the other to an event after t. Similarly, if we fix A, we get two different C's.

It is hard to disagree with these arguments. Nevertheless, it is also hard to accept fragility. Pressing the button as a fragile event is not fragile only in regard to time. That event can happen in infinitely many different space regions, depending on the position of my finger. I can press it in infinitely many different ways, using more or less power, under different angles, quickly or slowly, (not to mention that I can do it with my foot, my nose, or my book). Are all these events different and are the propositions referring to them all different? Is only one of the events a truthmaker for the proposition 'I press the button'? And what about our ordinary language practice – do we distinguish all these propositions in our everyday communication? It seems that extreme fragility is too problematic a notion[77].

It seems to me that what counts as the same proposition depends on the context. In his recent paper on causation Lewis got rid of the problem of fragility of events by using (implicitly) the notion of fragile propositions[78]. A similar solution might be attempted here. We can stop talking about propositions that antecedents and consequences stand for, and talk about sentences instead. A sentence would stand for a cluster of fragile propositions, i.e. it could stand for (possibly infinite) disjunction of fragile propositions. What is to be included in the cluster remains vague and depends on the context. A sentence would be true iff the disjunction is true. Lots of technical details remain to be worked out to develop this idea and see if it's any good, but I will not attempt to do it here. I will just add a few remarks. If there is a time gap between the antecedent and the consequent then the antecedent of CSO is never true and CSO is always vacuously true. This was not the intention of the theorists who proposed semantics that validate CSO. They wanted to express a certain feature of absolute similarity using CSO, namely that whenever the closest A-worlds are C-worlds, and the closest C-worlds are A-worlds, then these worlds are the same. They did not intend to state a claim that is vacuous because it never happens that the closest A-worlds are C-worlds, and at the same time the closest C-worlds are A-worlds. Besides that, these theorists often treat propositions as non-fragile (as everybody often does). Just remember all those counterexamples to contraposition:

---

[77] These problems with fragility of propositions are parallel to the problems of fragility of events, as described in Lewis's papers on causation. Cf. See Lewis 1973a, Lewis 1986b, Lewis 1986c (especially Postscript E on late preemption) and Lewis 2000.

[78] Lewis 2000.

$$\frac{A \to C}{\neg C \to \neg A}$$

When it is said that contraposition is not valid, it is not because the two A's (or the two C's), considered as fragile, cannot be the same. The claim that contraposition is invalid is not meant to be vacuously true.

To avoid the problem of fragility, we can attack CSO in a different way. First, we can try to make A, C, and D vague this way:

A_1 = I press the button somewhere around midnight

C_1 = The light is on somewhere around midnight.

D_1 = I hear the anthem somewhere around midnight.

If this is still not convincing enough, we can try something else. Conditionals involving causal relations and processes are likely to have a time gap between the antecedent and the consequent. So we can avoid those and try to make examples that involve institutional facts. That way we can have simultaneous antecedents and consequents. Gabbay proposed such an example[79]:

A_2 = I am elected president of the US.

C_2 = I am recalling the US troops from Asia.

D_2 = I am nicely dressed.

$A_2 \to C_2$, $C_2 \to A_2$, and $A_2 \to D_2$ might be true, but, Gabbay says, they do not imply $C_2 \to D_2$.

---

[79] Gabbay 1976 p. 190.

The counterexample might be made more persuasive if we found propositions such that not only is $C_2 \rightarrow D_2$ false, but $C_2 \rightarrow \neg D_2$ is true. Let us replace D_2 with D_3:

D_3 = Mom is happy

$A_2 \rightarrow D_3$ being true does not need much explanation. $C_2 \rightarrow \neg D_3$ is true, and hence $C_2 \rightarrow D_3$ is false, because my mom's political views make her strongly opposed to the US troops withdrawal.

There is still something wrong with the counterexample. I think that $C_2 \rightarrow A_2$ is either backtracking or for some other reason false. Recalling the troops would not make me an elected president. Rather, me being a president is what would make the antecedent true. By the antecedent being true I mean that I *successfully* recalled the troops. However, with the antecedent being true in this sense, the conditional is backtracking. If we understand the antecedent as my *unsuccessful* recalling, (which makes more sense, since I have no authority) then $C_2 \rightarrow A_2$ is false, and the following is true:

Had I recalled the troops, everybody would have laughed at me.

Let us therefore replace A_2 and C_2 with

A_3 = I have the authority to command the US troops

C_3 = The troops obey me to withdraw

Now $C_3 \rightarrow A_3$ is not backtracking because it is the troops' obedience that gives me the authority, or power to command, which I did not have before they obeyed. It would not be a legal authority, but it would be some kind of authority, since my 'commands' are being obeyed. Again $A_3 \rightarrow C_3$, $C_3 \rightarrow A_3$, and $A_3 \rightarrow D_3$ are true, but $C_3 \rightarrow D_3$ is not.

Let us see now how Gabbay and the defenders of CSO explain this counterexample. Gabbay's explanation might be derived from the following citations:

> [When asserting a conditional $A \to B$ the speaker wants to assert that] B follows from A (i.e. that $A \supset B$ holds) under 'certain' conditions. These conditions depend on the meaning of A and B and on the properties of the world in which $A \to B$ was uttered.
>
> For example, if I say, 'If I were the president I would have withdrawn from the east', I mean to say that the political situation being the same, B follows from A (i.e. $A \supset B$). ... So in order to falsify my statement, one has to present a possible world where both the political situation is the same and I am president but where I do not withdraw from the east. We do not care whether in that world a Mr. Smith has a beard or not, because this is not relevant to my statement.[80]

Gabbay's reasons for thinking that $C_2 \to D_2$ is false are stated very briefly:

> It is improper to say, if I were to recall the troops, I would have been elected president and therefore would have been nicely dressed. That makes $A_2 \to D_2$ context dependent on other sentences or on the entire conversation. When I utter $A_2 \to D_2$ I don't mention C_2 at all, and therefore its truth value depends on $g(A_2, D_2, i)$ alone.[81]

Thus Gabbay's explanation of $C_3 \to D_4$ (Had the troops obeyed me to withdraw, mom would be happy) being false might be this. Mom is strongly opposed to withdrawal, and would oppose it even if I ordered it. Me making the troops withdraw means also that I have the authority to command, but this fact is not relevant here. The selection function for $C_3 \to D_3$ depends only on the present situation and C_3 and D_3. A_3, which says that I have the authority, is irrelevant as well as a Mr. Smith's having beard is irrelevant in the citation above.

---

[80] Gabbay 1976 p. 187.
[81] Gabbay 1976 p. 190.

Defenders of CSO would say that A_3 is certainly relevant for $C_3 \rightarrow D_3$, because C_3 has as a consequence A_3 (i.e. $C_3 \rightarrow A_3$ is true as well), and A_3 has lots of influence on my mom's happiness (i.e. on D_3). It might easily be true that she would be still happy if I had the authority, even if I did something she finds foolish, like withdrawing the troops. Note that not only defenders of CSO might say that, but the defenders of antecedent-relativity as well would agree in this case, although they reject CSO.

Gabbay would say that it is 'improper' to involve A_3 in evaluating $C_3 \rightarrow D_3$ because when uttering $C_3 \rightarrow D_3$ I do not mention A_3 at all. That would make $C_3 \rightarrow D_3$ context dependent on other sentences or on the entire conversation. What Gabby had in mind is probably that the truth value of the conditional should depend on the ternary function $g(C_3, D_3, i)$, and not on some function $g(C_3, D_3, A_3, i)$, which has one more proposition as an argument. It is in accordance with Gricean tradition to say that the literal meaning and truth value of a conditional should not depend on what was previously said in the conversation. Logics for counterfactuals normally assume this either implicitly or explicitly.

However, CSO defenders would say that they do not mean that a function with four arguments is needed, but simply that A_3-worlds must be in the set $g(C_3, D_3, i)$, i.e. they should hold at the C_3-worlds from $g(C_3, D_3, i)$. Their remark would not be that some arguments should be added to Gabbay's function g, but that there already are too many arguments. They would attack the most distinctive feature of Gabbay's theory, namely that his function selects different worlds for counterfactuals that have the same antecedents but different consequents. They believe that if C→A and C→D are both true, C→A∧D must be also true. The ternary function does not allow that inference. Gabbay, they might say, went too far with his relativity.

However, this inference, Gabbay would say, is simply not valid. This is shown by the same counterexample to CSO. There are very few inferences that hold in Gabbay's semantics. 'This seems very little, but our examples give us no alternative.' – these are the very last words in Gabbay's paper[82].

I spent so much space discussing these counterexamples because CSO is an important formula, because they help me explain Gabbay's theory, and because they lead us to new issues, like backtracking and fragility, and to one more interesting idea that I will present now. Gabby raised the question of a context dependency of a kind that has never been discussed in the literature on counterfactuals until recently, namely the question whether our previous conversation should count as a context that determines the meaning and the truth value of a conditional. It might be that Gabbay was not quite happy with his counterexample, since right after the last citation above he said that 'of course' we can define semantics differently, so that the function g 'remembers' that A_3 was involved in conversation before, and $C_3 \rightarrow D_3$ is true iff $g(C_3, D_3, A_3, i) \subseteq [C_3 \supset D_3]$. This was, however, just a short remark. As far as I know, it had no influence on the development of logic for counterfactuals. Nevertheless, I think that the idea is very important, and I will discuss it in the next chapter.

The counterexample to CSO can still be improved by making the fourth conditional more obviously false. A_3 and C_3 remain the same, and instead of D_3 we have

---

[82] Gabbay op. cit.

D_4 I feel obliged to withdraw the troops.

Thus the final version of the counterexample goes as follows. It is true that:

$A_3 \rightarrow C_3$ If I had the authority to command the US troops, they would obey me to withdraw.

$C_3 \rightarrow A_3$ If the US troops obeyed me to withdraw, I would have the authority to command them.

$A_3 \rightarrow D_4$ If I had the authority to command the US troops, I would feel obliged to withdraw them.

But it is false that:

$C_3 \rightarrow D_4$ If the US troops obeyed me to withdraw, I would (still) feel obliged to withdraw them.

Moreover, the opposite is true:

$C_3 \rightarrow \neg D_4$ If the US troops obeyed me to withdraw, I would not (any more) feel obliged to withdraw them.

Would this version of the counterexample make a difference? I believe that the switch to this version goes in Gabbay's favour, but the CSO-defenders are not beaten yet. In an absolute similarity semantics we imagine a situation (i.e. worlds) where the antecedents of the above conditionals hold. These are worlds where I have the authority to command and where the troops obey me. Since the third conditional in the example ($A_3 \rightarrow D_4$) is true as well, it is also true in those worlds that I feel obliged to withdraw the troops. If now the fourth conditional ($C_3 \rightarrow D_4$) is false, this means that I do not feel obliged to withdraw the troops at some of those worlds, the very same worlds where I do feel obliged to withdraw. The worlds are accordingly impossible (D_4 and $\neg D_4$ both hold there). But the closest worlds where a contingent antecedent holds cannot be impossible.

Gabbay would say that these are not the same worlds where D_4 and $\neg D_4$ hold. D_4 holds at some C_3 worlds where A_3 does not hold, and $\neg D_4$ holds at some A_3 worlds where C_3 does not hold. Thus impossible worlds are not involved.

Let us consider the above four conditionals from the point of view of the metalinguistic theory. Throughout this thesis I have assumed that the basic idea of the metalinguistic theory is right: $A \rightarrow C$ is true iff the argument from A, and some background truths cotenable with A, to C is valid. The assumption is perfectly safe. You cannot reject it, since it is so vague that it can fit with whatever you believe about conditionals. It's just a basic idea, not a theory. Thus we can say that whatever opposes this idea is wrong. CSO defenders might say that Gabbay's and my counterexamples oppose it. Let us say that it is true that the fourth conditional $C_3 \rightarrow D_4$ is false. It makes no sense to feel obliged to withdraw the troops when they are already withdrawing. However, this implies that the third conditional $A_3 \rightarrow D_4$ is also false. All the propositions involved – A_3, C_3, and D_4 – pertain to the same moment *t*. If C_3 is true whenever A_3 is (and that is what the second conditional asserts), that is, if the troops obey me to withdraw when I have the authority, then it makes no sense for me to feel obliged to withdraw the troops when I have the authority because they are already withdrawing. In other words, the third and the fourth conditional are either both true or both false. While asserting that the third conditional is true I made an assumption that the troops are not (yet) obeying me at *t* (i.e. $\neg C_3$) even if I have the authority at *t*. Only under the assumption that $\neg C_3$ would still be true even if A_3 were true can the third conditional be asserted. $\neg C_3$ is therefore one of the background propositions for the third conditional. However, that contradicts the second premise: $C_3 \rightarrow A_3$. If C_3 is a background proposition for the third conditional, than it has to be cotenable with A_3, i.e. $\neg(A_3 \rightarrow \neg\neg C_3)$, which is equivalent to $\neg(A_3 \rightarrow C_3)$, and that is the negation of the second premise. Thus my counterexample rests on a contradiction. And the analoguous argument can be made against Gabbay's original version of the counterexample.

I hope I have been clear enough in expressing the intuition against Gabbay's and my view on CSO. Can we respond? The argument against us is strong, but not conclusive (indeed, hardly any claim about conditionals is conclusive). We can say, first, all the worse for the metalinguistic theory. This answer requires a clear explanation of what is wrong with the metalinguistic theory. I would rather go some other way, and use my notion of conditionals going and not 'going together'. I haven't explained what it means that two or more conditionals do 'go together', but I believe that I made it clear that there are examples of conditionals that cannot 'go together'. The above argument claims that Gabbay and I fell into contradiction, because we claim explicitly that one conditional is true (the second conditional $C_3 \rightarrow A_3$), and implicitly we claim that the same conditional is false (because we also assert the third conditional $A_3 \rightarrow D_4$, which assumes the falsity of the second conditional). However, before accusing us of inconsistency one has to show that the second and the third conditional 'go together'. This cannot be just assumed, because it is not trivial. Before deciding who is right we have to explain 'going together'. I will attempt to do that in the next chapter. But even before that Gabbay might give an answer to the above remark.

We should keep in mind that we do not have a 'real' definition of cotenability (at least in this thesis). All we have is a circular definition of cotenability in terms of counterfactuals. We decide whether a proposition is cotenable with another proposition according to some intuitions about certain counterfactual. But we have different intuitions about counterfactuals. The above argument accusing Gabbay and me of inconsistency uses intuitions that are not Gabbay's. They could be Lewis's, for example. The truth value of a conditional and the selection of important worlds depend on both the antecedent and the consequent, says Gabbay. Not so for Lewis – he would say that the choice of important worlds does not depend on the consequent. It would not be fair to accuse Gabbay of inconsistency because he assigns certain truth values to some conditionals that are different from truth values that Lewis's semantics assigns to them. When evaluating A→C some propositions are cotenable with A and they hold at the important worlds that the function g selects. But what those propositions are depends on both A and C. When we evaluate A→D some other propositions may be cotenable with A. Therefore there might be a proposition B cotenable with A in the context of evaluating A→C, while in the context of evaluating A→D, B is not cotenable with A. Thus the accusation of inconsistency is not valid.

It is time to see who is the winner in this debate for and against CSO. I proclaim the game tied. Obviously we are just running in circles. Both sides use the very same example to support their positions, and both have their way of defending themselves from the opponent, but in a way that the opponent cannot accept. They seem to be speaking in different languages. Thus we did not advance in showing whether CSO should or should not hold. Still there is a conclusion that I think may be derived from this tied result. There are two conflicting intuitions about the counterexample. CSO supporters and Gabbay can each explain only one intuition. Thus they both score, but no theory is good enough. A satisfactory theory should be able to explain the opposing intuition, and to explain why it was easy for the opponent to make the mistake. This is not a case where one side is clearly wrong. We feel that both sides grab a piece of truth. Claiming that one side is completely right amounts to saying that the other is irrational, or at least not too smart. This, however, is too strong a claim for me to accept. In the next chapter I will try again to resolve this dispute.

I have already rejected absolute theories in previous chapters, and here is my main argument against Gabbay's theory. Although I have a lot of sympathy for Gabbay's approach, his theory is clearly unacceptable. This is because there are almost no rules of inference involving conditionals that are valid in G. Nothing holds there. Rational people use in ordinary language and science lots of rules that do not hold in G. Ergo, Gabbay's theory does not describe our ordinary language practice. QED. Antecedent relativity is not much better in that regard. That notion gives us also a very weak logic, which cannot say much about the way we actually reason. Absolute theories are stronger than these two relative theories, and are better in that respect. However, absolute theories can also be said to be weak, and for that reason they were quit a surprise in the late 60's and early 70's when they were discovered. Stalnaker and Lewis presented counterexamples to transitivity and contraposition, rules that were sacred until then. Their systems are weaker than, at that time, one would suppose conditional logic to be. Today, however, it is Warmbrōd's theory that is considered heretic, because it validates transitivity and contraposition. The third kind of relativity that I will discuss is context-relativity It is

Warmbrōd's device. It is exactly the opposite of Gabbay's relativity, in the sense that while almost no rule holds in Gabbay's logic, Warmbrōd's theory validates far more rules than any other logic for counterfactuals. According to the opinion of most people interested in conditionals, Warmbrōd went too far, just as Gabbay did, but in the opposite direction.

6.4 Warmbrōd's context-relativity

A model for Warmbrōd's semantics is a quadruple $< i, \mathrm{I}, \mathrm{R}, [\,] >$ where I is a (non-empty) set of possible worlds, R a binary reflexive (accessibility) relation on I, $i \in \mathrm{I}$, and $[\,]$ is a function which assigns to each proposition ϕ a subset $[\phi]$ of I (all the worlds from I where ϕ is true). ϕ is said to be true under $< i, \mathrm{I}, \mathrm{R}, [\,] >$ iff $i \in [\phi]$.

> $\phi \rightarrow \psi$ is true under $< i, \mathrm{I}, \mathrm{R}, [\,] >$ iff for each $j \in \mathrm{I}$ such that $i\mathrm{R}j$ holds, $\phi \supset \psi$ is true under $< i, \mathrm{I}, \mathrm{R}, [\,] >$.

The similarity between Warmbrōd's conditional and the strict implication from modal logic is obvious: a conditional is true iff the corresponding material implication is true at all accessible worlds. Thus $\phi \rightarrow \psi$ is $\Box(\phi \supset \psi)$, where the box has the same formal properties as the necessity operator from Feys-von Wright modal theory T (whether it's important for Warmbrōd that the modal system is T and not stronger, I don't know; but the system cannot be weaker, since R is reflexive).

Obviously, then, Warmbrōd's counterfactuals obey the same rules of inference that hold for strict implication, including transitivity, contraposition and strengthening of antecedents, which were famously rejected by the first possible worlds theories of Lewis and Stalnaker. However, for various reasons, some people thought that Stalnaker's and

Lewis systems were still too strong. Plausible counterexamples were proposed for several rules. Objections to CEM, CS, CV, and CSO have already been mentioned in previous chapters. Nute argued against SEA (Substitution of Equivalent Antecedents)[83], and McKay and van Inwagen objected to SDA (Simplification of Disjunctive Antecedents)[84]

SEA $$\frac{\phi \equiv \psi}{(\phi \rightarrow \chi) \equiv (\psi \rightarrow \chi)}$$

SDA $$\frac{(\phi \vee \psi) \rightarrow \chi}{\phi \rightarrow \chi}$$

Counterexamples to most of these rules were known before 1981 when Warmbrōd's paper was published (he discussed especially the last two). Faced with the counterexamples, we can either accept a logic *far* weaker than Stalnaker's or Lewis's (which is presumably unacceptable), or face the problem of explaining away the counterexamples. Since we must address that problem anyway, it seems appropriate, Warmbrōd said, to reconsider even the initial rejection of transitivity and contraposition[85]. His solution to the problems with SEA and SDA enables him at the same time to explain away the counterexamples to transitivity and contraposition.

The solution consists in supplementing the semantics defined above with a pragmatic theory that explains the accessibility relation R. The other three elements from the model, I, *i*, and [], can be regarded as fixed, while R varies from one occasion of speech to another. Thus R is always relative to a particular piece of discourse, and it is determined primarily by a conditional appearing early in the discourse. The first conditional is evaluated according to what Warmbrōd called the *standard interpretation*, which is so called probably because it resembles the most popular Stalnaker-Lewis approach. $\phi \rightarrow \psi$ is true iff ψ holds at the closest ϕ-worlds, assuming that ϕ is not *absurd*. These truth conditions presuppose the limit assumption, which Warmbrōd defends in another paper[86]. Warmbrōd lets us decide on our own what we want to call 'absurd'. An absurd antecedent can be logically impossible, or be true only at the very distant worlds, or something like that. Warmbrōd thinks that we have no natural way of interpreting conditionals with such antecedents, and his theory leaves them uninterpreted.

---

[83] Nute 1975b.
[84] McKay and van Inwagen 1977.
[85] Warmbrōd 1981. p. 279.

Thus the relation R is determined by the standard interpretation of a conditional early in the conversation. R picks up the closest worlds at which the first antecedent is true, and is then held constant even to evaluate other conditionals that we want to consider in the same corpus. A new corpus requires a new accessibility relation. How do we know when one corpus ends and a new one begins? We keep the same R as long as new conditionals obey the *normality condition*, which says that any new antecedent must hold at some accessible worlds. The normality condition is generalized from examples like this:

6.9 If Aunt Brachia had a baby, she would be an unwed mother

Aunt Brachia's spinsterhood is a background assumption in our conversation and for that reason 6.9 is true. This conversational assumption is reflected, as Warmbrōd says, in the accessibility relation that results from the standard interpretation: Aunt Brachia is single at all the closest worlds where she has a baby. Imagine now that somebody introduces the conditional:

If Aunt Brachia got married, she would cease to be a spinster

---

[86] Warmbrōd 1982.

The new antecedent conflicts with our assumption that Aunt Brachia was single (i.e. it breaks the normality rule). If we keep the old accessibility relation, the new conditional is trivialized. To avoid the trivialization, we consider the new conditional as the beginning of a new conversation.

Warmbrōd's theory can be expressed in more precise term as follows. Let D be a body of discourse, *i* the actual world, and '$\leq$' the similarity relation.

W_1 (Standard interpretation) If $\phi \to \psi$ is the first conditional that occurs in D, iRj iff: $j \in [\phi]$ and for any $k \in [\phi]$, $j \leq_i k$.[87]

W_2 (Normality condition) If $\phi \to \psi$ occurs in D, then for some $j \in [\phi]$, iRj.

W_3 (Truth conditions) $\phi \to \psi$ is true at i iff for every iRj, $j \in [\phi \supset \psi]$.

Let us now see how the theory can be used to explain away the counterexamples to transitivity. 6.9 and 6.10 by transitivity entail 6.11, but 6.11 is obviously false, even if we assume that the premises are true.

6.10 If Aunt Brachia were married, she would have a baby

6.11 If Aunt Brachia were married, she would be an unwed mother.

The accessibility relation is not normal (in the sense of W_2) throughout the argument; otherwise, 6.10 would be trivialized. Thus the argument must be regarded as two bodies of discourse. The two premises cannot be evaluated in the same model. Warmbrōd's pragmatics offers a similar explanation of counterexamples to contraposition:

6.12 If Aunt Brachia had a baby, she would not get married.

---

[87] Warmbrōd didn't exactly say that the accessible worlds are determined only by the first conditional. His exact words are that R is determined early in D. My guess is that he would allow other non-conditional sentences to influence R as well (such as: 'Let us assume p.'; it is then natural to suppose that p would hold at all the accessible worlds, even though it not the antecedent of a conditional occurring in D). If we focus only on conditionals, then I believe that W_1 is what Warmbrōd had in mind.

6.13 If Aunt Brachia got married, she would not have a baby.

Again the antecedent of the second conditional breaks the normality rule.

Warmbrōd's theory is in accordance with the fact that we use transitivity and contraposition very often. That fact certainly needs an explanation, because according to other theories our usage of these rules is simply irrational. Lewis emphasized that $(A \rightarrow B) \wedge (A \wedge B \rightarrow C) \supset (A \rightarrow C)$ is valid in his semantics[88], and that rule might be offered as what we 'really mean' when using transitivity. However, I know of no attempt to save contraposition. Whoever reads these lines is very likely to believe that Henkin's proof of completeness[89] is right. However, Henkin never proved that if $\models A$ then $\vdash A$. He proved that if not $\vdash A$, then not $\models A$. Often instances of transitivity and contraposition look logically necessary, and a satisfactory conditional logic should explain that.

Nute rejects Warmbrōd's theory because he thinks that a safer approach would be better. The problem is to provide an account of the difference between the situations in which the rules are reliable and those in which they are not. Warmbrōd's strategy is to consider the rules always reliable and then to provide a way of falsifying the premises in 'unhappy cases'. A better approach would be to consider these rules invalid and then look for those features of context that sometimes allow us to use them with impunity. 'It is probably better to occasionally overlook a good argument than it is to embrace a bad one. Or to put a bit differently, it is better to force the argument to bear the burden of proof rather than to consider it sound until proven unsound.'[90]

---

[88] Lewis 1973 p. 35.
[89] Henkin 1949.
[90] Nute and Cross 2002 p. 24

More can be said about the good and the bad sides of Warmbrōd's theory, and of other theories that I have mentioned so far. I leave that for the next chapter. All these theories solve some problems that other theories cannot, and they all have their own problems. We will see which of those problems could be solved. For the end of this chapter I would like to add some more comments on the metalinguistic theory.

6.5 Relative theories and the metalinguistic theory

In the previous chapters I have compared each of the presented possible worlds semantics with what I called Nute's and Lewis's versions of the standard interpretations of the metalinguistic theory (Nute's UI and Lewis's UI). Now I will do the same with the relative theories. To distinguish different versions of the metalinguistic theory, I find it useful to use the notion of *relevant* background propositions. That notion is explained case by case. For example, the relevant background propositions for the conditionals 4.1 (Had I struck m, it would have lit) are $B_1 - B_4$ (which say that m is dry, well made, etc.). Nute's UI requires not only the relevant propositions, but also *all* cotenable truths to be among the background propositions. Lewis's UI is intended to include the relevant propositions and beside them *some but not all* truths cotenable with the antecedent; only such cotenable truths are included whose negation is less possible than the antecedent. We saw in Chapters 3 and 5 that Nute's UI validates CEM, but is not equivalent to Stalnaker's theory because it allows a countermodel to CSO. Lewis's UI does not validate CEM, but is still not equivalent to Lewis's own semantics, because the latter fails to satisfy 5.0 (p. 64 above), and thus cannot be related to any version of the metalinguistic theory.

We can notice some similarities between Nute's version and the version of the antecedent-relative minimal change semantics that I described above, namely the version that satisfies 3.0 (see p. 22 and p. 72). We saw that both validate CEM and neither validates CSO. In Mårtensson's version of an antecedent-relative semantics CV:

$$((\phi\rightarrow\psi)\wedge\neg(\phi\rightarrow\neg\chi))\supset((\phi\wedge\chi)\rightarrow\psi)$$

does not hold. Let us see the status of that formula in the metalinguistic theory. Suppose that χ and $\neg\chi$ are both cotenable with ϕ, i.e. both $\phi\rightarrow\neg\chi$ and $\phi\rightarrow\chi$ are false. Suppose further that $\neg\chi$ is true. Then $\neg\chi$ has to be a background proposition for any conditional with ϕ as an antecedent. Suppose that $\neg\chi$ is a *relevant* background proposition for $\phi\rightarrow\psi$, i.e. ψ is a logical consequence of $\neg\chi$ and ϕ (and possibly some other cotenable truths). In that case the argument

$$\text{6.14} \qquad \frac{\phi, \ldots, \neg\chi, \ldots}{\psi}$$

is valid, i.e. $\phi\rightarrow\psi$ is true, and

$$\text{6.15} \qquad \frac{\phi\wedge\chi, \ldots}{\psi}$$

is invalid, i.e. $\phi\wedge\chi\rightarrow\psi$ is false. This makes CV invalid. Note that the proof assumes that there can be a proposition and its negation both cotenable with another proposition. This is possible in Lewis's UI but not in Nute's UI, because the latter validates CEM. Thus CV fails in Lewis's UI, but it holds in Nute's UI. Proof of the latter: If χ is cotenable with ϕ, i.e. $\neg(\phi\rightarrow\neg\chi)$, as it is stated in the antecedent of CV, then by CEM $\phi\rightarrow\chi$ and $\phi\rightarrow\neg\neg\chi$ are true, which means that $\neg\chi$ is not cotenable with ϕ. Then we have two possibilities: χ

is true or not. If it is true, then χ is among the background propositions for $\phi \rightarrow \psi$. The antecedent of CV in that case says that

$$6.16 \qquad \frac{\phi, \ldots, \chi, \ldots}{\psi}$$

is valid, which entails that 6.15 (the consequent of CV) is also valid. If χ is false, then $\neg\chi$ cannot be among the background propositions for $\phi \rightarrow \psi$, since it is not cotenable with ϕ. In that case the antecedent of CV says that

$$6.17 \qquad \frac{\phi, \ldots}{\psi}$$

is valid, which again entails the validity of 6.15. Thus CV holds in Nute's version.

Nevertheless, the rejection of CV is not an essential feature of the antecedent-relative version of minimal change theories[91]. CV can be rejected for other reasons[92], but we do not have to reject it in order to accept antecedent-relativity. From Fact 2 above I concluded that each truth that would not be false if A were true holds at the closest A-worlds, i.e.

$$6.18 \qquad (C \wedge \neg(A \rightarrow \neg C)) \supset (A \rightarrow C),$$

which I take to be intuitively an obvious result of the mixture of the ideas of antecedent-relativity and minimal change and 3.0. We already encountered this formula in Chapter 3 where, following Bennett, we called it PF*. We proved there that 6.18 entails CEM (theorem 1 on p. 28). It is not difficult to see that Nute's UI and the antecedent-relative minimal change semantics where 6.18 holds express the same idea in different terms.

---

[91] Note that Mårtensson's theory is not exactly of the type of antecedent-relative theory that I describe here. His theory is more complicated, with a similarity defined in terms of causal notions, and it does *not* validate CEM. Besides making these few notes, I will not discuss this theory any more.

[92] See Pollock 1981 p. 254f.

6.18 obviously follows from Nute's UI. On the other hand, we can derive Nute's UI from the truth conditions for '→' as defined in the antecedent relative theory: A→C is true iff C holds at the closest A-worlds, i.e. at the closest A-world, since we proved that there can be only one. That A-world is determined by A and every truth whose negation is not needed to make A true. And that is the world determined by A and all the truths cotenable with it. A→C is true, i.e. C holds in that A-world, if and only if the argument from A and all cotenable truths to C is valid.

Lewis's UI is weaker than Nute's, because it does not validate CEM nor CV. CSO does not hold there either, which can be seen from the same countermodel we made within Nute's version. CV and CSO are axioms for Lewis's favourite system VC, so the difference between his own theory and what I called Lewis's UI is obvious. It is then clear why Lewis had to invent a radically different notion of cotenability in order to relate his semantics to the metalinguistic theory.

It is therefore an important question what is to be included among the background propositions, because different answers give us different formal properties of '→'. Let us now see what happens if we exclude from the background propositions everything that is unnecessary and keep only the relevant propositions. Let us once again consider the conditionals 4.1 (Had I struck m, it would have lit = A→C). The relevant background propositions are $B_1 - B_4$ (m is dry, well made, etc.). Of course, what the relevant propositions are depends on both A and C. Thus this version of the metalinguistic theory reminds one of the antecedent-and-consequent relativity. Further, if A→C is true, then the argument from A and $B_1 - B_4$ to C is valid. It is valid iff C holds at all A-and-$B_1 - B_4$-worlds. These worlds resemble the closest A-worlds insofar as the four B's hold at them, but otherwise they can defer from them to any degree whatsoever. This feature reminds one of maximal change. Thus this version of the metalinguistic theory has both essential features of Gabbay's theory – antecedent-and-consequent relativity and maximal change.

In Chapter 3 (p. 31) I cited Bennett's words that we should include among the background propositions other propositions besides the relevant ones, for two reasons. First, it is too much trouble to keep them out, and, second, they do no harm ('Irrelevance is harmless'). The second reason turned out to be false. I do not agree with the first one either. We do have a tool that can help us get rid of irrelevant background propositions. We can say that A→C is true iff there is a valid argument of the form {A, ...}⊢ C, just as it is in all the metalinguistic theories, but we can change the notion of deduction. The turnstile '⊢' can be defined not in classical logic, but in relevance logic. This kind of theory would be very similar to Gabbay's theory, especially in the standard cases of counterfactuals with contingent antecedents, but it would not be the same. For example, not all conditionals with an impossible antecedent would be true, as it is the case in Gabbay's theory. I don't know if Gabbay would accept this modification.

Since the version of the metalinguistic theory where all the background propositions are relevant is like Gabbay's theory, it has to be very weak. I will not go into details about which rules fail to hold here, but I will mention CSO again because of an important moral that will come out. CSO obviously doesn't hold here. The countermodel that we used within other versions of the metalinguistic theory will work in this case as well, but I would like to add one more. Let A→C and C→A and A→D be true, and let A, C, and D be contingent and false (A and C must not be logically equivalent). ¬C is therefore true. The conjunction of A and ¬C (and possibly some other truths cotenable with A) must have some logical consequences (different from those each of the two conjuncts has by itself). Suppose that D is one of them. In that case A→D is true, but C→D cannot be true, contrary to what CSO predicts. This is the general pattern I have used to make each of the counterexamples to CSO in this chapter. One might say that this trick is a bit dirty, because ¬C is one of the background propositions for A→D, and hence should be cotenable with A, while the second premise A→C implies that ¬C is not cotenable with A. Does it mean that this type of theory ignores Goodman's (very plausible) requirement that the background propositions must be cotenable with the antecedent? I do not think so, but there is an important novelty that should be emphasized. We should remember that we do not have a real definition of cotenability, but only a circular definition in terms of counterfactuals. Now, counterfactuals are antecedent-and-consequent relative, and so must be cotenability as well. What is cotenable with the antecedent of a conditional A→C is relative to both A and C. If it were relative only to A, that would be antecedent(only)-relativity. That way we can have ¬C among the background propositions for A→D, even though we also want to assert A→C.

Thus we saw in this section that Nute's UI is a metalinguistic counterpart of the antecedent-relative possible worlds semantics that satisfies 3.0. Lewis's UI corresponds to a different kind of antecedent-relative semantics. The metalinguistic theory that allows only the relevant cotenable truths to be among the background propositions would be a counterpart of Gabbay's theory. What is left to be done is to compare Warmbrōd's theory to a version of the metalinguistic theory. I leave that for the next chapter.

7. A context-relative theory of counterfactuals

We saw that each of the possible worlds semantics we discussed has some problems. The topic of this chapter is to examine if the problems can be solved or avoided. One of the problems is that the theories sometimes seem to be too strong, sometimes too weak. I will first try to offer an intuitive explanation of why the relative theories, except Warmbrōd's, are weaker than the absolute theories.

Relative semantics have their selection functions of the form f(A, *i*) or g(A, C, *i*). f is an antecedent-relative, and g is an antecedent-and-consequent relative selection function. Within these semantics we can define the notions of a model and of entailment or validity of arguments. A model is usually defined as a triple or a quadruple which includes the selection function as one of its elements. An argument is valid if there is no model under which the premises are true and the conclusion false. Of course, if we evaluate in a model an argument that involves more than one counterfactual, all counterfactuals are evaluated according to the same selection function, since the function is determined by the model. We can use models to evaluate sets of propositions as well. These sets can be maximal. When a model makes true every proposition from a maximal possible set, we can say that the set 'describes a world', our own, or some other possible world. Obviously, when we make a model for a set that includes more than one conditional, they are all evaluated with the same selection function.

These are all very simple things and very common for logical systems. Without these features we would hardly call a system 'logical'. Nevertheless, let us imagine some strange 'logical' systems that do not have these simple features. Imagine an antecedent-relative theory where for each antecedent we use a different selection function, say f*.

Thus $f^*_A(i)$ is the function that selects the closest A-worlds relative to the world i, and is different from the function $f^*_B(i)$ or $f^*_C(i)$ etc. However, $f^*_A(i)$ selects the same worlds as the 'normal' function $f(A, i)$, as well as $f^*_B(i) = f(B, i)$, $f^*_C(i) = f(C, i)$, etc. Similarly, let g^* be the strange version of the 'normal' function g. g^* is different for any pair of antecedents and consequents. Thus, for example, $g^*_{A,C}(i)$, which is used to evaluate $A \rightarrow C$, is a function different from $g^*_{A,B}(i)$or $g^*_{B,D}(i)$, which are used to evaluate $A \rightarrow B$ and $B \rightarrow D$, even though $g^*_{A,C}(i) = g(A, C, i)$, $g^*_{A,B}(i) = g(A, B, i)$, $g^*_{B,D}(i) = g(B, D, i)$, etc. Let the notions of a model and entailment for the 'starred' logic be the same as for the 'normal' logic.

What is the result? The starred logics are limited in evaluating sets of propositions that include counterfactuals – models can be made only for those counterfactuals that can be evaluated with the same selection function. Thus the starred antecedent-relative logic can evaluate in one model only counterfactuals with the same antecedent, while the starred antecedent-and-consequent relative logic needs a different model for each counterfactual. Accordingly, the arguments can be valid only if the premises and the conclusion can be evaluated in the same model, i.e. using one starred selection function. The starred logics cannot 'describe the world', i.e. no maximal models can be made.

Nevertheless, since the starred and the 'normal' function select the same worlds in each particular case, the set of theorems and valid rules in the starred logics is exactly the same as in their corresponding 'normal' versions. This story about the starred logics has the purpose to offer an intuitive explanation of why the relative theories are weaker than the absolute ones: it is easy to make a counterexample to a rule of inference if the premises and the conclusion are evaluated in different models. The antecedent-relative theories are weaker because they treat validity *as if* the counterfactuals with different antecedents were evaluated in different models. The antecedent-and-consequent-relative theories are the weakest because they treat validity *as if* the counterfactuals with different antecedents and consequents were evaluated in different models, that is, as if each counterfactual were evaluated by itself. In the absolute theories all counterfactuals 'go together', i.e. all counterfactuals can be evaluated in the same model. In the relative theories counterfactuals go together if they have the same antecedent, or if they have the same both the antecedent and the consequent (i.e. they go together only with themselves).

Another purpose of the story about the starred logics is to emphasize that, unlike other relative theories discussed above, Warmbrōd's logic does not have its 'normal' version. In Warmbrōd's logic models are relative to a body of discourse and we cannot make maximal models. Warmbrōd's semantics cannot give us a description of the world, and it can evaluate in the same model only conditionals that satisfy his normality condition (W_2). In Gabbay's theory, for example, we can have a consistent set containing every possible proposition (including counterfactuals) or its negation. Or, to say the same thing in terms of starred-Gabbay's theory, we can describe the (actual) world with a set containing all the counterfactuals that are true each in its own model (relative to the actual world). Within Warmbrōd's theory it makes no sense to make a set of all counterfactuals that are true in some body of discourse, since there is no single normality condition that they all could obey. That would clearly be an inconsistent set and not the best description of our world.

When introducing each of the relative theories in the previous chapter I tried to relate them to the problem of conditionals 'going together'. A single selection function cannot make an ordering of worlds that yields the right truth value for all counterfactuals. That is the ambition of the absolute theories. I believe I showed that they fail to achieve this goal, because some counterfactuals do not go together. When we adjust the ordering of worlds to get the right truth value for some counterfactual, some other counterfactual will have the wrong truth valued according to that ordering. We can regard the antecedent relative theories, as described in this thesis, as an attempt to solve the question – which conditionals do go together? However, these theories turned out to be 'not relative enough'. The second proposal, offered by the antecedent-and-consequent relativity, turned out to be 'too relative', i.e. that no two conditionals go together. We need a solution somewhere in between. Warmbrōd's normality condition W_2 is such a solution. However, I will propose a different one.

7.1 A modification of Warmbrōd's theory

Here is the proposal meant to replace W_2. It is formulated in terms of the metalinguistic theory, and uses the notion of background propositions and Goodman's notion of cotenability. We can say that

7.1 some counterfactuals can *go together* only if each of the background propositions for any of the counterfactuals is cotenable with each of their antecedents.

On one hand, this proposal solves the problem of absolute theories. The general pattern for making counterexamples to these theories (Chapter 5) was as follows. We need three truths of the form $\phi \rightarrow \psi$, $\neg\chi_1 \rightarrow \phi$, and $[\neg\chi_1] <_i [\chi_1 \wedge \phi]$, where χ_1 is one of the background propositions that make $\phi \rightarrow \psi$ true. (For example, had I struck m, it would have lit; had it been wet, I would have struck it). Since χ_1 is a background proposition for

$\phi \rightarrow \psi$, and χ_1 is not cotenable with the antecedent $\neg\chi_1$, $\phi \rightarrow \psi$ and $\neg\chi_1 \rightarrow \phi$ do not go together. On the other hand, by allowing some conditionals to go together, the proposal solves the problem of Gabbay's theory, which does not allow any two conditionals to go together. This enables us to define validity so that some rules would be valid (as opposed to Gabbay's theory where none is valid).

7.1 can be used as a basis for a possible worlds semantics. The basic idea is the same old one: we need a selection function that picks up the *important* antecedent worlds. Gabbay's ternary selection function is, I believe, the best choice. The truth conditions for a counterfactual are then defined in Gabbay's way. After that we leave Gabby and go Warmbrōd's way. We give a restricted notion of validity of rules of inference, which works as the usual notion of validity, but only if the premises and the conclusion go together, as defined in 7.1. Instead of Warmbrōd's notion of 'body of discourse' and his normality condition, I use the notion of *context* and 7.1. Conditionals that go together belong to the same context, and context is defined as a set containing all the background propositions of all the counterfactuals we are considering. To distinguish the ordinary language notion of context from the one I just defined, the letter will be put under quotes ('context').

The similarity of my proposal to Warmbrōd's theory is obvious. What good do we get from the differences?

First, the notion of 'context' as a set of background propositions can help us answer an important question: When we have two propositions A and C, what proposition is expressed by A→C? This is not a problem, for example, for Lewis. A→C is the proposition defined as a set of worlds where A→C holds. In Warmbrōd's theory, on the other hand, it is not clear how to answer the question, because what A→C really says is relative to a body of discourse. When asking what proposition is expressed by A→C, we want to know what the conditional says *absolutely*, not relatively to something. To deal with this question I introduce the notion of *underlying logic*. Counterfactuals are context-dependent. The underlying logic has to be context-free. We have lots of context-free logics at our disposal, because that is usually how propositional logics are done (classical, intuitionistic, modal, relevant etc etc propositional logic). Let '⇒' be the implication from the underlying logic. '→' behaves the same way as '⇒' behaves in the underlining logic when the members of 'context' are taken as assumptions. Thus if we want to define propositions as sets of worlds, we can say that the proposition A→C is the set of worlds where A⇒C and all the members of the 'context' hold. Or, if we want to use the metalinguistic formulation, A→C says that there is a relation of entailment between the proposition A and all the members of 'context' on one side, and the proposition C on the other side; the notion of entailment is defined within the underlying logic. How do we choose the underlying logic? That depends on one's logical taste. Apparently Warmbrōd's taste would lead him to choose the propositional modal logic T. I would choose relevance logic. [93]

The second benefit we get from modifying Warmbrōd's theory, I believe, is this. Warmbrōd's relying on the notion of similarity leads him to some unnecessary complications, unnecessary because they have a lot to do with similarity and very little with conditionals. If our body of discourse begins with A→C and later we want to assert B→D, according to the normality condition W_2 (If $\phi \rightarrow \psi$ occurs in D, then for some $j \in [\phi]$, jRi.), some B worlds must hold at some closest A-worlds. This means that B has to be equally or more similar to actuality than A. In the general case, each new antecedent has to be equally or more close to actuality than the first antecedent. However, there is no intuitive justification for this one-way restriction. We might want to introduce a conditional whose antecedent represents a slightly more remote possibility than the first antecedent. This is certainly what we often do in ordinary language. Does it mean that we then start a new body of discourse? Even if the new conditional does not mess up with the truth values of previous conditionals? Does it mean that in ordinary language we can never consider such conditionals in a single argument? The normality condition requires an affirmative answer to these questions. This intuitively wrong result is baggage that Warmbrōd's theory carries as an inheritance from the similarity theories. I will now try to explain why that happens.

---

[93] I will not discuss here the reasons for choosing one or the other underlying logic. Not only because 'de gustibus non est disputandum', but because this would be a big digression for a thesis on counterfactuals, and because I would not be able to give some essentially new defense of relevant logic, other than the standard defense offered, for example, by Anderson and Belnap 1975, Anderson, Belnap and Dunn 1992, Stephen Read 1988, or Mares and Meyer 2001.

One of the purposes of the normality condition is to ensure that the first antecedent does not exclude as a possibility some later antecedent. How do we say in conditional logic that A does not exclude B? There are two (equivalent) ways to say that in Lewis's logic: A◊→B (if A were the case, B might be the case), or we can say that in Goodman's sense B is cotenable with A: $\neg(A \rightarrow \neg B)$. (A◊→B is equivalent to $\neg(A \rightarrow \neg B)$ by definition). For these formulae to be true B must hold at some closest A worlds. This is the same as Warmbrōd's normality requirement if A is the antecedent of the first and B the antecedent of some later conditional from the same body of discourse. However, might-conditionals are unnecessarily complicated if defined in terms of similarity. If

A◊→B then B must be equally or more close to actuality than A. If B◊→A also holds, then A and B must be equally close to actuality.

7.2 If I were in Neverland, I might be rich

because, as you know, it's a rich country and nobody there is paid better than philosophers. Also

7.3 If I were rich, I might be in Neverland,

because the folks there are corrupt, and I could bribe them to hire me even though I am not capable of writing anything better than this thesis. Thus we have both A◊→B and B◊→A. Suppose further that I am rich (A) but I am not in Neverland (¬B). Certainly then A is closer to actuality than B, contrary to what Lewis's semantics predicts. The problem is that the similarity relation or a selection function based on similarity restricts the choice of the antecedent-words only to those that are the closest. That way we lose some relevant worlds. For that reason I prefer to talk about *important*, rather than similar worlds. For A not to exclude B, or for A◊→B, or for B to be cotenable with A, it is enough that B holds at least at some A-worlds that are in certain respects the same as our world. What makes these worlds 'important' is being 'the same in certain respects', and not 'being the closest'. Beside these 'certain respects', the important A-worlds can differ from our world to any degree whatsoever. That gives us the right truth values for 7.2 and 7.3. You have probably recognized that I am using here the idea of 'maximal change', which is the distinctive feature of Gabbay's theory. Beside other good aspects of that theory mentioned in the previous chapter, this is one of my most important reasons to modify Warmbrōd's theory by replacing the elements of Lewis's semantics with the elements of Gabbay's theory.

To emphasize the bad aspects of the normality condition, I give another example.

7.4 If Otto had come, it would have been a lively party

7.5 If Otto and Ana had come it would have been be a lively party

therefore

7.6 If Otto had come, or if they had come together, it would have been a lively party

Otto and Ana are even more fun when they are together. However, the worlds where Ana comes are more remote than those where Otto comes without her. Thus 7.5 does not satisfy the normality condition. The argument is of the form

$$7.7 \qquad \frac{\begin{array}{c} A \to C \\ B \to C \end{array}}{A \vee B \to C}$$

This argument is valid in Warmbrōd's semantics. However, because of the normality condition, the 7.4-7.6 argument, which is an instance of 7.7, cannot be evaluated in Warmbrōd's semantics. Thus we lose an instance of an intuitively valid inference. My point is that there are inferences of the form that Warmbrōd's semantics considers valid, but which are not captured by Warmbrōd's semantics. Although I cannot prove that my modification of Warmbrōd's theory does not have the same kind of problem, it is certainly an improvement in that regard, because for the 7.1 requirement it is irrelevant whether the second antecedent is closer than the first one, and it considers the 7.4-7.6 argument valid.

Here is what I have done so far in this chapter. I do not believe that world ordering semantics can work because no ordering is good for all counterfactuals. Thus we need some restriction when evaluating counterfactuals, that is, we should evaluate together only those conditionals that do not mess with each other. Putting the restriction that only counterfactuals with the same antecedent go together (as in the strange antecedent-relative version) is not good, because it is in some regards too strong and too weak in others. It is too strong because some counterfactuals with different antecedents do *not* mess with each other, and too weak because all counterfactuals with a given antecedent do mess up when put together (section 6.1 above). Antecedent-and-consequent relativity is too strong a restriction, which can be shown by pointing to any pair of counterfactuals that do not mess with each other. Warmbrōd's normality condition is another restriction. Its good side is that it is weaker than both above restrictions, in that it allows conditionals with different antecedents to be evaluated together. It is still too strong in some cases, like the 7.4-7.6 argument, where it does not allow conditionals that do not mess up to be evaluated together. The 7.1 restriction keeps the good sides of Warmbrōd's restriction and avoids the bad side.

On the other hand, I do find Gabbay's arguments convincing when he says that the selection of important worlds should depend on the consequent as well. This gives us good truth conditions for counterfactuals evaluated in isolation.

As I said above, Warmbrōd's semantics does not have its non-starred version. This means two things: it cannot give us a model that would describe the world, and it is not clear what proposition is expressed by A→C, even when it is clear what propositions are expressed by A and C. This problem can be removed by noting the role played by the underlying logic. In Warmbrōd's case this is the modal logic T. If A→C is at the beginning of a body of discourse D, the set of worlds that characterizes D is the set of the closest A-worlds. There is a set X of propositions such that all members of X are true at the closest A-worlds, and nowhere else are they all true. A→C then behaves the same way as the strict implication A⇒C behaves in the system T under the assumption of all the members of X. Thus the question of finding the proposition expressed by A→C is to be answered in the usual way, as it is done in modal logic, since we found a way to translate A→C into the modal logic T. Once we know what proposition is expressed by a counterfactual, we can describe the world, i.e. we can make maximal models, simply because maximal models can be made in the system T. In the similar way the same problems are solved in the case of my modification of Warmbrōd's theory.

Nevertheless, something still remains strange. The two features of logical systems – describing the world, i.e. making maximal models, and defining validity of arguments as not having false conclusion and true premises in any of these models – are something normally defined in the same system. One of my main conclusions in this section is that this cannot be done in conditional logic. There must be a division of labour: we need a *pragmatic* part (a 'pragmatic semantics'), as it is in Warmbrōd's case the system T together with W_1 and W_2 (standard interpretation and the normality condition), which has the task of telling us which rules are valid, and we need an underlying logic, as it is in Warmbrōd's case the system T without W1 and W2, which is context-free, and which has the task of describing the world and expressing the propositions of the sentences from the pragmatic part. Conditional logic must have a pragmatic part that distinguishes good from bad reasoning, and it must have a context-free part that does the work of the underlying logic. Counterfactuals are highly context-sensitive. Ordinary language, and hence ordinary language reasoning and argumentation, are context-sensitive. No context-free logic can describe or explain our ordinary language practice. The pragmatic part will have a task of capturing our ordinary language reasoning by defining validity of rules of inference. The 'stable' context-free underlying part will have other purposes, as described above.

Stalnaker's and Lewis's semantics, as is well known, are also context-dependent. Stalnaker's selection function and Lewis's similarity relation are not fixed once and for always by the central world, but depend on the context of utterance as well. Different similarity measures are appropriate in different situations. However, once we decide what similarity is appropriate for our present purposes, according to their theories all counterfactuals are to be evaluated using that similarity relation. I have tried to show that this cannot be done, because, while that similarity measure will be appropriate for some counterfactuals, there will always be other counterfactuals that will enforce a *shift of context*, and thus require a different similarity measure. An example of a shift of context is a shift of 'context' (which I defined as a set of background propositions), i.e. when the previous 'context' or previous antecedent is not cotenable with the antecedent of a new counterfactual. The shift in 'context' means that what counts as important in our selection of worlds has changed, and different worlds must now be selected as important. Hence the shift of 'context' requires a change of the ordering of worlds, i.e. new similarity measures.

My original plan when I started writing this thesis was to propose a theory of counterfactuals that would be a modification of Warmbrōd's theory as described above. W_1 (standard interpretation) that borrows from Stalnaker and Lewis would be replaced by elements borrowed from Gabbay; W_2 (normality condition) would be replaced by 7.1, i.e by the definition of 'going together'; and the underlying T would be replaced by relevance logic. I also planned to give an alternative to 7.1, which would make transitivity invalid, so that we could have definitions of contexts where transitivity is safe, and those where it is not. However, reading von Fintel made me change the plan. Lots of details remain to be worked out, but I will not do them. I realized that the theory would anyway not be good enough, and that one crucial step must be added. But before I explain that I have first to present von Fintel's theory.

7.2 von Fintel's dynamic semantics

> A primary goal of research in the semantics/pragmatics interface is to investigate the division of labor between the truth-conditional component of the meaning of an expression and other factors of a more pragmatic nature. One favorite strategy, associated foremost with Grice[94] is to keep to a rather austere semantics and to derive the overall meaning of an utterance by predictable additional inferences, called "implicatures", which are seen as based on certain principles of rational and purposeful interaction. In this chapter, I will explore a different way in which the truth-conditional component is complemented in context.[95]

Within the Gricean tradition we can say that, if an expression α in a context *c* expresses the proposition p, an appropriate way to capture this in a semantic system is "to attribute to α a context-dependent meaning that maps *c* to p"[96], i.e.

7.8 $[\alpha]^c = p$

---

[94] Grice 1967, 1989

[95] von Fintel 2001 p. 123.

[96] Ibid.

Von Fintel proposes a different analysis, and that is the crucial step that I accept here. The other analysis attributes to α a meaning that has two aspects. First, α alters the initial context *c* to a new context *c'* (7.9 a.). Then *c'* maps α to the proposition p (7.9 b.) in a systematic, and, as von Fintel says, simpler way than under analysis 7.8.

7.9 a. $c|\alpha| = c'$

b. $[\alpha]^{c'} = p$

($[\alpha]$ denotes α with respect to the contextual parameter *c* and is a set of worlds; $|\alpha|$ is the *context change potential* of α and is a function from contexts to contexts; in accordance with the practice in dynamic logic, the function is written to the right of its argument.)

Which of the two analyses is right is to be determined empirically. Here is the evidence von Fintel offers in favour of 7.9.

7.10 If the USA threw its weapons into the sea tomorrow, there would be war; but if all the nuclear powers threw their weapons into the sea tomorrow, there would be peace.

This is one of Lewis's arguments that counterfactuals are not strict implications and that strengthening of antecedent:

$$\frac{A \rightarrow C}{A \wedge B \rightarrow C}$$

is not a valid rule of inference. (You can notice that the second conditional in 7.10 does not satisfy Warmbrōd's normality condition, nor my 7.1, and therefore requires a different 'context' than the first conditional.) Von Fintel mentions 7.10 as an example of a sequence of conditionals for which Lewis would say that the context remains the same throughout the sequence, so that both conditionals can be evaluated in the same system of spheres. Similarly, Dorothy Edgington argues against strengthening of antecedent[97]

> [A] piece of masonry falls from the cornice of a building, narrowly missing a worker. The foreman says: 'If you had been standing a foot to the left, you would have been killed; but if you had (also) been wearing your hard hat, you would have been alright.'

Von Fintel agrees that Edgington quite correctly says that the building foreman's remarks constitute "a single, pointful piece of discourse", and he adds that one can read them as a "shrewd way of putting the suggestion" that the worker should wear a hard hat at all times. But then von Fintel draws our attention to the following example by Irene Heim[98]:

> 7.11 ? If all the nuclear powers threw their weapons into the sea tomorrow, there would be peace; but if the USA threw its weapons into the sea tomorrow, there would be war.

> In 7.11, says von Fintel, the two counterfactuals claimed to be consistent by Lewis are reversed in order and the sequence does not work as before. The reason seems intuitively clear: once we consider as contextually relevant worlds where all nuclear powers abandon their weapons, we cannot ignore them when considering what would happen if the USA disarmed itself. We seem to be in need of an account that *keeps track* of what possibilities have been considered and does not allow succeeding counterfactuals to ignore those possibilities. An

---

[97] Edgington 1995 pp 252-3. Edgington's argument can be understood as a counterexample to Warmbrŏd's theory. However, she does not refer to Warmbrŏd, but to Crispin Wright 1983 (Warmbrŏd's paper is two years earlier than Wright's). Wright makes a short comment where he proposes a context-relative theory of counterfactuals similar to Warmbrŏd's. Apparently neither Wright nor Edgington were aware of Warmbrŏd's paper.

[98] From an MIT seminar in the spring of 1994.

> account according to which the context remains constant throughout these examples would not expect a contrast between the two orders.[99]

We can admit that 7.10, as well as Edgington's example, constitute "a single, pointful piece of discourse". That, however, does not mean that we cannot have a context change within a single pointful discourse.

> Lewis deliberately put this example in the form of a single run-on sentence, with the counterfactuals conjoined by semicolons and *but*. This was meant to ensure that the context stays constant throughout, an assumption that in our more dynamic days seems rather naïve.
>
> ...
>
> The proper diagnosis would seem to be that over the course of 7.10, the set of worlds quantified over properly expands, but that over the course of 7.11, it cannot shrink. This asymmetry is unexpected if one maintains there is no context change.[100]

Another unexpected thing for the 'static' approach is that there are examples where we can appropriately say that the initial conditional is 'no longer' true. For example:

7.12 A: If the USA threw its weapons into the sea tomorrow, there would be war; but if all the nuclear powers threw their weapons into the sea tomorrow, there would be peace.

---

[99] von Fintel, op cit. pp. 130-131. I insert the italic, which does not appear in the original text, wanting to remind the reader that we have already encountered a somewhat similar idea in the previous chapter, when we mentioned Gabbay's selection function that 'remembers' what was previously said in conversation. The idea is in accordance with von Fintel's dynamic theory of meaning (7.9). Gabbay, however, spent only two sentences on the idea.

[100] von Fintel, ibid. p. 131.

B: But that means that if the USA threw its weapons, there wouldn't *necessarily* be war

B': But that means that if the USA threw its weapons, there *might not* be war

What B' says directly contradicts the first A's conditional.

If we go back to the simpler antecedent, the domain of quantification should shrink back to the closest worlds where just the USA disarms, ignoring the far-fetched worlds where all the nuclear powers become meek. But this does not seem to happen.[101]

In his defense of transitivity Warmbrōd pointed to the fact that the counterexamples lose their strength when we change the order of the premises:

7.13 If Aunt Brachia had a baby, she would be an unwed mother
If Aunt Brachia were married, she would have a baby
Therefore, if Aunt Brachia were married, she would be an unwed mother.

7.14 If Aunt Brachia were married, she would have a baby
? If Aunt Brachia had a baby, she would be an unwed mother
Therefore, if Aunt Brachia were married, she would be an unwed mother.

Something different is going on in 7.13 and in 7.14, but a static approach cannot explain that. Warmbrōd's solution and defence of transitivity includes showing that one of the premises is false. Von Fintel, who also considers this phenomenon in another counterexample to transitivity, would say that in 7.14 we are not at all tempted to admit both premises as true. The natural way of reading the second premise would be to take into account the worlds where Aunt Bracia had a baby *and* where she was married, since the latter is already introduced by the first premise.

---

[101] Ibid.

Thus we can say that the dynamics of meaning is a detectable phenomenon and that von Fintel's 7.9 makes sense. Another argument in favour of 7.9 is that von Fintel's semantics based on it is useful, as we will see now when we turn to the notion of validity of arguments. An argument from $\phi_1, ..., \phi_n$ to ψ is (dynamically) valid iff

$$7.15 \quad [\phi_1]^c \cap [\phi_2]^{c|\phi_1|} \cap ... \cap [\phi_n]^{c|\phi_1| ... |\phi_{n-1}|} \subseteq [\psi]^{c|\phi_1| ... |\phi_n|}$$

A sequence is dynamically consistent iff

$$7.16 \quad [\phi_1]^c \cap [\phi_2]^{c|\phi_1|} \cap ... \cap [\phi_n]^{c|\phi_1| ... |\phi_{n-1}|} \neq \varnothing$$ [102]

Here is what the definitions mean when the propositions stand for counterfactuals. The first counterfactual in a conversation is to be evaluated 'standardly', i.e. it is true iff the consequent holds at the closest antecedent worlds (the same as in Warmbrōd's theory). The closest antecedent worlds are determined by a selection function *f*. *f* is the dynamic element of the semantics and it is sensitive to context. In Lewis's theory *f* is context sensitive as well, but it does not depend on our previous conversation. In von Fintel's case, the previous conversation is a part of the context that influences *f*. He has a convenient name for the selection function – it is a 'modal horizon' that expands to include worlds where each new antecedent holds. More precisely:

> If a conditional is accepted as an assertion, the context will first be changed to expand the modal horizon if the antecedent wasn't already considered a relevant possibility. Then, the conditional will be interpreted in the new context. What we would like to do then is to assign the counterfactual $\phi \rightarrow \psi$ a context change

---

[102] Ibid p. 142

> potential, a function from contexts to contexts that changes the context so as to add the antecedent to the modal horizon. The proposition expressed by the conditional is then computed with respect to the already updated context.[103]
>
> Assume that initially, *f* is trivial in that it assigns to each world *w* only $\{w\}$. Now, $\phi \rightarrow \psi$ is offered. *f* needs to be expanded. Apart from *w*, we need to have in *f(w)* the closest ϕ-worlds and all additional non-ϕ-worlds that are closer to *w* than the closest ϕ-worlds. The conditional now claims that all of the ϕ-worlds in *f(w)* are ψ-worlds. [104]

Von Fintel's counterfactual resembles Warmbrōd's since both are strict implications. While Warmbrōd's accessibility relation is static, von Fintel's 'modal horizon' can change during the conversation. Note that the modal horizon can only expand. It never shrinks. This is not so because von Fintel thinks that modal horizon never shrinks in ordinary speech. On the contrary, he thinks that it does, but so far he does not have a technical solution that could deal with this phenomenon.[105]

Let us now reexamine the examples given in this section to apply von Fintel's notions and see how they work. The difference between 7.10 and 7.11 is in their dynamic consistency:

7.10 If the USA threw its weapons into the sea tomorrow, there would be war; but if all the nuclear powers threw their weapons into the sea tomorrow, there would be peace.

---

[103] Ibid. 9. 138

[104] Ibid. p. 139

[105] Ibid. 139ff. Shrinking will be considered in the next section.

7.11 ? If all the nuclear powers threw their weapons into the sea tomorrow, there would be peace; but if the USA threw its weapons into the sea tomorrow, there would be war.

> The [7.10] sequence is dynamically consistent because we can start with a context whose modal horizon is just wide enough to include those ϕ-worlds that are ψ-worlds; this horizon is then widened by the second sentence, which may well be true if all of the closest $\phi\wedge\chi$-worlds are non-ψ-worlds. The Heim sequence [7.11] is dynamically inconsistent, because we have no automatic mechanism that would allow the horizon to shrink between the addition of the first sentence and the assessment of the second sentence. As a result, the first sentence makes the claim that the $\phi\wedge\chi$-worlds in the set of accessible worlds are all non-ψ-worlds while the second sentence makes the claim that all ϕ-worlds in the very same set of accessible worlds are ψ-worlds: a straightforward contradiction.[106]

The difference between the versions of the counterexample to transitivity 7.13 and 7.14 is explained by the first being dynamically invalid and the second valid. In general, transitivity is dynamically invalid if the premises are ordered in one way: $\psi\to\chi$, $\phi\to\psi$, therefore $\phi\to\chi$. It is valid if the premises are ordered in the other way: $\phi\to\psi$, $\psi\to\chi$, therefore $\phi\to\chi$. Besides this exception, most of the monotonic inference patterns are dynamically invalid. In that regard von Fintel's theory is more similar to the standard theory than to Warmbrōd's.

---

[106] Ibid. 142.

However, von Fintel introduces one more notion of entailment, according to which an argument from $\phi_1, \ldots, \phi_n$ to ψ is valid iff

7.17 for all contexts c such that $c = c|\phi 1| \ldots |\phi n|$,

$$[\phi_1]^c \cap [\phi_2]^{c|\phi 1|} \cap \ldots \cap [\phi_n]^{c|\phi 1| \ldots |\phi n\text{-}1|} \subseteq [\psi]^{c|\phi 1| \ldots |\phi n|}$$ [107]

Unlike the notion of dynamic validity 7.15, 7.17 is supposed to capture the notion of validity of 'logical arguments'. Von Fintel did not explain what exactly he meant by the notion of a logical argument, but he said that in logical argumentation we are committed to a stable context. "In classical logic, it is considered an imperative that in the assessment of arguments the context remain stable".[108] Of course, 7.17 considers monotonic rules valid.

7.3 On four aspects of context-dependency

Let us leave von Fintel's theory for a while. I would like now to turn my attention to the notion of context. I am interested in features of context that affect truth values of counterfactuals. There are at least four factors that influence context in the way that affects truth values of conditionals: (1) the previous conversation, (2) the context of utterance not related to the content of our previous conversation, but determined by the state of affairs at the time and place where the communication happens, and/or time and place to which the communication refers, (3) the meaning and the context-dependence of the antecedents and the consequents by themselves, especially their fragility, and (4) the principle of charity. These factors are usually not independent. They can influence each other.

---

[107] Ibid. p.143.

[108] Ibid. p. 141.

Context dependence of counterfactuals in the literature on counterfactuals usually assumes only (2). It is captured in the metalinguistic theory by the set of background propositions, and in possible worlds semantics by the selection function or similarity ordering. (1) is a relatively new topic, introduced by von Fintel's dynamic theory. (4) is not discussed often, but it has its role in von Fintel's theory. I haven't heard of anybody discussing (3) in the literature on conditionals, but Lewis's work on causation and events is extremely useful for the topic[109]. Let us consider some examples to see how the four factors work.

7.18 Had I put the thermometer t into boiling water, it would have shown 100°

is true here, but false in the Himalayas. When I talked about this example with my colleagues, some of them agreed, and others first reacted by denying that the conditional was the same in these two cases, because the antecedents were different. If uttered once here, the other time in the Himalayas, *propositions* expressed by the two antecedents would refer to two different events. Therefore we do not have the same conditional uttered in the two different contexts, but two different conditionals (i.e. two different propositions that are expressed by the same conditional sentence uttered in two different contexts). Let it be so for now, and let us consider again the conditional uttered here (not in the Himalayas). What if I put the thermometer into water a tenth part of the second later? What if I put it a bit to the left? Would that affect the antecedent? And if t showed 100°, even though I put t a bit to the left, would we say that this is *not* a confirmation of the same conditional 7.18, but of another one? Obviously in usual situations we do not care about the *exact* spatio-temporal region where my putting t into water happens. Some changes are permitted. But how big are the changes that are permitted? Some of my colleagues would say that if it happened on Himalayas, it would be a different proposition. Thus what we need is to put (discover?) a border somewhere between here and Himalayas, and that would enable us to tell exactly...

---

[109] See Lewis 1973a, Lewis 1986b, Lewis 1986c (especially Postscript E on late preemption) and Lewis 2000.

That won't work, obviously. I don't think that it makes sense to decide about such borders once and for always. Where the border is depends on the context. In different situations and different conversations we would assume different borders. Of course, the borders are often very vague, as it is the case in the traditional examples of vagueness, about the heaps, for example. So I think that we should allow that the two utterances of 7.18 *could* have the same antecedents and the same consequents in *some* contexts. There is a clear sense in which we can understand the antecedent uttered here and uttered on Himalayas as standing for the same proposition: it is the same thermometer t, the same water, the same me, and, if you want, the same *putting* (the same force under the same angle). So both events – putting here and putting on Himalayas – could be truthmakers for the same sentence 'I put t into boiling water'.

Why did some of my colleagues say that the two antecedents were different? Referring just to the change of place is *ad hoc* and not a good *general* explanation: do we want to say that the consequents are different as well? I don't think so. Change of place seems to be irrelevant there. There was a Volkswagen Rabbit, said Quine, with 'Gavagai' written on the plates[110]. Is that still true when the car is moving? Or do we have a continuum of *different* propositions of the form '"Gavagai" is written on the plates (here/now)'? Obviously the change of place is irrelevant here, but why does it matter for the antecedent in the thermometer case?

---

[110] Three Indeterminacies

Imagine a club of classic drama lovers. The only member who did not go yesterday night to see 'Electra' was Tim, who did not expect the main actress to be capable of a good performance. Today the club meets and Tim's friends try to convince him that he made a mistake, because

7.19 Everybody who went to the theatre had a great time, and

7.20 Had Tim come he would have had a great time too.

To show that he was not wrong Tim has to find at least one person who went to the theatre and did not have fun. That would falsify the first statement (7.19) and cast doubt on the counterfactual 7.20.

7.21 Jim was at the theatre early this morning cleaning the washrooms and did not have fun.

Obviously nobody would say that 7.21 casts any doubt to 7.19 and 7.20, but let us spell out the reasons for which this is so. This sentence does not falsify those two for two reasons. It breaks two rules that we follow in successful conversations. The first one is about the context-dependence of the quantifiers. 'Everybody' is determined by the previous conversation in the club and means either 'all the members' or 'people from the audience during yesterday night's play'. Therefore it does not include Jim. The other is that although the proposition 7.19 'Everybody who went to the theatre had a great time' is vague with respect to time, it is not vague enough to include people who went to the theatre before or after yesterday night. The meaning of the antecedent of 7.20

7.22 Tim comes

has certain fragility with respect to both space and time. It means that Tim comes to the theatre (not somewhere else, and not anywhere in the theatre, but to the place where he could watch the show), which is determined by 7.19 and previous conversation, and is sensitive with respect to time in the same way as 7.19. The principle of charity seems to be a good explanation of how we manage to understand each other. In particular, it seems to be a good explanation of how we determine the meaning of 7.19, 7.20 and 7.22 – we interpret them in the way that makes 7.19 and 7.20 true.

It was long ago that logicians realized that before interpreting quantified sentences we need to specify a domain of individuals we are talking about. What the domain is depends on context. It has therefore been a long time ago that we have allowed pragmatics (in this sense) to be part of truth conditions, and a factor in determining what proposition corresponds to a sentence. Context (in this sense) is thus considered an inseparable part of formal semantics for predicate logic, and this fact is well known (although not always expressed in this way that emphasizes the role of pragmatics). Therefore, we would do nothing essentially new if we allowed pragmatics to be a factor in formal semantics for counterfactuals: context should tell us what propositions stand for our conditional sentences, and what propositions stand for our antecedents and consequents. This is exactly what we do in classical predicate logic anyway. We allow the meaning of 7.19 to be determined by the context, but the meaning of 7.20 is determined by the same context.

After we first encountered the problem of fragility (section 6.3), I suggested that it should be solved by defining propositions as sets of fragile propositions. Thus the proposition 7.22 is a set containing propositions that say that Tim comes at 9pm, or a ½ second after, from the northern or from the southern side, dressed in a blue or black suit, etc, etc. 7.22 is vague in many respects, not only time and place. For example, 7.23 Had Tim come naked, he would have a great time

is not what the club members meant. Thus there is no naked Tim in the set for 7.22. Let big Greek letters stand for propositions and small ones for fragile propositions.

7.24 $\Phi \rightarrow \Psi$ is true iff $(\forall \phi \in \Phi)(\exists \psi \in \Psi)(\phi \rightarrow \psi)$ is true

What the antecedent and the consequent mean, i.e. what fragile members of the antecedent and the consequent are, obviously, has influence on the truth value of the conditional. However, there is influence in the other direction as well. The latter influence is explained by the principle of charity. If there is room for different interpretations, we often interpret the meaning of the antecedent and the consequent in such a way that would make the conditional true. For example, we interpret the meaning of the antecedent and the consequent of 7.20 in such a way that 7.21 does not oppose 7.20, or in such a way that the antecedent of 7.23 is not a subset of the antecedent of 7.20.

Since I first started learning philosophy and logic, I used to believe that clear ordinary language indicative sentences, which we have no problem understanding, express something precise and absolute that we call propositions. 7.25 below is an example of a sentence expressing such a ‘proposition’. Now I believe that ordinary language sentences that express something precise and absolute are very rare. Sentences expressing fragile propositions would be of that kind. There are also sentences that do not express fragile propositions, but still express something precise and absolute. These could be defined as sets of fragile propositions, but all aspects of their fragility would have exact borders. Most sentences, however, express something that can be defined as a set of fragile propositions, but the aspects of their fragility have very imprecise borders. Thus I suggest that we distinguish three kinds of propositions: (i) *fragile* propositions, whose truthmaker cannot be altered in any way; (ii) *exact* propositions, which have more than one truthmaker, and which can be defined as sets of fragile propositions; those sets have precise borders in the sense that for any fragile proposition we know whether it belongs to the set or not; whenever a fragile proposition from the set is true, the exact proposition is true; (iii) *vague* propositions, which are the most common, and which are defined as sets without precise borders. The exact place of the borders is highly context-sensitive, and in great majority of cases it remains vague (the degree of the vagueness again being highly context-sensitive). Let us consider an example from one of the standard textbooks[111]:

> To know the meaning of a sentence is to know its truth-conditions. If I say to you
>
> 7.25 There is a bag of potatoes in my pantry
>
> you may not know whether what I have said was true. What you do know, however, is what the world would have to be like for it to be true. There has to be a bag of potatoes in my pantry. The truth of 7.25 can come about in ever so many ways. The bag may be paper of plastic, big or small. It may be sitting on the floor or hiding behind a basket of onions on the shelf. The potatoes may come from Idaho or northern Maine. There may even be more than a single bag. Change the situation as you please. As long as there is a bag of potatoes in my pantry, sentence 7.25 is true.

7.25 is a typical example of a sentence that I used to believe to express a 'proposition', something stable and absolute. It now seems to be of the second kind I mentioned above, namely, what is expressed by 7.25 could be defined as a set of fragile propositions, but the set would be well defined – for any fragile proposition we could say with no problem whether it belongs to the set or not. A sentence could hardly be more clear than 7.25 is! However, as I said, I do not believe that any more. 7.25 expresses a proposition of the third kind – a set of fragile propositions with vague borders that we are ready to revise whenever the context shifts in some relevant way.

---

[111] Heim and Kratzer 1998 p. 1.

My pantry was not built with a bag of potatoes in it. Thus there was a time when there was not a bag of potatoes in my pantry. This is, of course, completely irrelevant for the truth value of 7.25, because 7.25 does not pertain to that time in the past when the pantry was built. What time does it pertain to, though? Two answers might be proposed. First, 7.25 can pertain to a *period* of time that includes the present moment, or the moment when 7.25 has been pronounced. Second, 7.25 may express a fragile proposition, i.e. it pertains only to one *moment*. The first answer assumes what my point is, namely that 7.25 is of the third kind, a set of fragile propositions with vague borders that can easily change as the context shifts. It is not of the second kind, because 7.25 does not say, for example, that there was a bag in my pantry from Wednesday 3:43 pm. till Friday noon. Thus whatever period of time 7.25 pertains to can be determined only vaguely. Since 7.25 is given out of any context of conversation, it is very vague. If we considered it as a part of some conversation, it could be less vague, or it could be even reduced to the second kind.

The first answer (that 7.25 pertains to a period of time) is right, because the second answer (that 7.25 expresses a fragile proposition that pertains not to a period but to an instance) leads us to all kinds of ridiculous questions. For example, what is the instance it pertains to? It takes some time to pronounce 7.25. It might pertain to the moment when I started pronouncing it, to the middle, or to the end of my pronunciation. Or does it pertain to the moment when you hear it (in which case the *meaning* of 7.25 depends on the distance, the speed of the sound, the temperature of the air, etc.)? Let us say that the most 'natural' answer would be that 7.25 pertains to the moment when I finish pronouncing the sentence. The problem that appears now is that at the end of the sentence I may add something that in no way influences the truth value. For example, I *proudly* say

7.26 There is a bag of potatoes in my pantry, baby!

Had I pronounced 7.26 instead of 7.25, I would have finished pronouncing the sentence a bit later; but we feel that 7.25 is true if and only if 7.26 is. Thus we can say that the moment the fragile proposition pertains to is the one when I finished pronouncing the part of the sentence that contributes to the truth value, i.e. the end of 'pantry'. But there are more problems:

7.27 There is a bag of potatoes in my pantry, and I mean it!

7.27 is a conjunction. What moments do the conjuncts pertain to? To the same moment when the whole sentence was pronounced, or does each conjunct have its own moment? None of the answers is in accordance with our standard inference patte[illegible]s for conjunction. If they pertain to different moments, then the conjunction does not follow from the conjuncts (the bag can be stolen while I am pronouncing the second conjunct). If they pertain to the same moment, what moment could that be? If it is the moment at the end of the whole sentence, then the first conjunct does not follow from the conjunction. If it's the moment at the end of the first conjunct, then the second conjunct does not follow from the conjunction. All this leads me to the conclusion that 7.25 does not express a fragile proposition. It is, therefore, a vague proposition.

Back to the thermometer example. We can now explain the sense in which some of my colleagues were right when claiming that the two utterances of 7.18 have different antecedents. *Knowing* enough physics, they interpreted the antecedent as more fragile in order to make the conditional true. The conditional about the thermometer was used as an example, taken in isolation and not as a part of some conversation. It was not clear then how to interpret the antecedent, and we cannot blame them for interpreting it as more fragile. But we cannot expect that the antecedent *must* be interpreted that way. In some contexts it will be appropriate to interpret it as more fragile, in some less. (Note that, had my colleagues not *known* enough physics, they would probably not have felt the need to interpret the antecedent as fragile. This is where an *epistemic* aspect comes into the semantics of counterfactuals, which are usually considered non-epistemic!)

Thus we saw that the four mentioned features of context can interact: following the principle of charity (4) we tend to interpret what is said by a counterfactual in a way that makes the counterfactual true, if there is room left for such an interpretation. Thus (4) influences fragility (3) of the antecedent and the consequent. Propositions can be fragile in an infinite number of ways. For example, my putting t into water can be fragile with respect to time, place, force, speed, angle of putting etc. The importance of each of these respects varies, so in different contexts some respects can become irrelevant, others can become important. (4) can influence (3) by determining which respects of fragility are important (namely, those that have influence on the truth value of the conditional). (3) in turn influences (2). (2) is a description of the state of affairs that holds in the relevant spatio-temporal region. The fragility of the propositions involved can expand or shrink the relevant spatio-temporal region, and that is how (3) influences (2). As I said above, by (4) we interpret a counterfactual in a way that makes it true, if there is room left for such an interpretation. What is it that determines if there is room left or not? In an initial context, or if a counterfactual is considered in isolation, possible interpretations are to some degree determined by the assumed meaning of the antecedent and the meaning of the consequent (separately) (3), which in turn are to some degree determined by the state of affairs in the region where the counterfactual was uttered (2). If the counterfactual was uttered later in a conversation, then the previous conversation (1) has influence on (2) and (3). Thus whether we can, following (4), interpret a counterfactual in a way that makes it true, depends on (2) and (3), which in turn depend on (1). If such an interpretation is possible, that may insert new influence on (2) and (3), and so on.

Sometimes, (1), (2), (3), and (4) can resolve all the vagueness of a conditional and its antecedent and consequent, but that happens only rarely. Usually some vagueness remains, but smooth conversation can go on. This is the most common situation in ordinary speech. We don't have many opportunities to use exact propositions, and there are even less opportunities to use fragile propositions. To borrow an example that John Perry[112] used in a somewhat similar context: 'It's raining'. Nothing easier than understanding this sentence, although it is highly context dependent and it is not possible to remove its vagueness. There is no way to define precise truth conditions for it, since that definition would have to specify some region to which the sentence pertains. But any precise borders of that region would be arbitrary.

---

[112] Perry 1986.

On the other hand, (1), (2), (3), and (4) might not be enough to reduce vagueness to a degree necessary for successful communication, which leaves room for misunderstandings. This was the case in the thermometer example 7.18.

It can also happen that the source of misunderstanding is a conflict between some of (1), (2), (3), and (4). An example of that would be a case where (put in von Fintel's terminology) the modal horizon shrinks. By (1) we would expect the modal horizon (i.e. (2)) to expand or remain the same, as von Fintel explained. If, at some point during a conversation, we introduce an assumption that conflicts with a previous assumption or antecedent, the principle of charity (4) can prevail over the influence of (1) and lead us to accept the new assumption and remove from the modal horizon the worlds where the old conflicting assumption holds. However, it is not always clear which one should prevail, (1) or (4).

In section 6.3 we left the questions raised by the counterexamples to CSO unanswered. Now we are in a better position to answer them. The first counterexample below involves the backtracking problem. In the second one A and C are simultaneous, so backtracking is avoided.

6.12 If I pressed the button, the light would be on. (A→C)

6.13 If the light were on, I would press the button. (C→A)

6.14 If I pressed the button, I would hear the anthem. (A→D)

But not

6.15 If the light were on, I would hear the anthem. (C→D)

$A_3 \rightarrow C_3$ If I had the authority to command the US troops, they would obey me to withdraw.

$C_3 \rightarrow A_3$ If the US troops obeyed me to withdraw, I would have the authority to command them.

$A_3 \rightarrow D_4$ If I had the authority to command the US troops, I would feel obliged to withdraw them.

But it is false that:

$C_3 \rightarrow D_4$ If the US troops obeyed me to withdraw, I would (still) feel obliged to withdraw them.

What does von Fintel's semantics say about the second counterexample? It says that the counterexample fails because the third premise is false. The first premise brings into the modal horizon the worlds where I have the authority. The second premise brings the worlds where the troops obey. The third has no influence on the horizon, but its consequent D_4 is false at the closest antecedent-worlds, since C_3 holds at those worlds (I don't feel obliged to withdraw the troops that are already withdrawing, the same as I don't feel obliged to pay an already paid debt). Thus the set of premises is not dynamically consistent. The set can be dynamically consistent if we change the order of the premises, for example if we put the third premise at the beginning. In that case von Fintel's semantics predicts that the conclusion is true as well. Whatever the order is, von Fintel's semantics says that the counterexample fails.

The aspect (1), previous conversation, is, of course, essential for von Fintel's estimation of the counterexample. As we said in section 6.3, (1) is exactly what Gabbay complains about. The third premise (in the order given above) is false for von Fintel because the first two are true. But even von Fintel would accept the third premise if it was considered in isolation, or if it was at the beginning of the conversation. Previous conversation should not be taken into account, said Gabbay, therefore the third premise should be considered true, and the counterexample works. Can we accept what Gabbay said?

I do not think that Gabbay's reasons should be accepted *generally.* Von Fintel gave us good reasons to consider the meaning of a counterfactual as dependent on the previous conversation. So we should not reject the whole dynamic approach. However, I think that there is something going on in this counterexample that might support Gabbay's position. Namely, it seems that the previous conversation should not be taken into account because the modal horizon should shrink. As we said above right before turning to the counterexamples to CSO, in the cases of shrinking we have a conflict between (1) and (4) (influence of previous conversation and charity). (4) leads us to accept the third premise once it was asserted, but the third premise assumes $\neg C_3$ (that the troops are not (yet) obeying me). However, according to (1), C_3 is already in the modal horizon, since it is the antecedent of the previous (second) premise. Therefore, if we want to accept the third premise, the modal horizon should shrink (which in this case is the same as going back to the initial context, as Gabbay requires).

Von Fintel considers the problem of shrinking[113]. He gives examples of the sequences of the form: {...; A→C; ...; But, of course, A would never happen; ...}. His explanation seems to be that the horizon shrinks in order to avoid a contradiction. I think that we can accept this as a general principle – the modal horizon should contract rather than allow contradictions. I don't know if there are other cases that require shrinking. Since it does not look likely, let us assume that shrinking occurs only to avoid contradictions. Von Fintel does not mention any examples where a conditional forces us to shrink the horizon. I have an impression that he thinks that conditionals cannot do that, i.e. it seems that he thinks that conditionals can only expand the horizon or leave it as it was, if the new antecedent is already in the horizon.

---

[113] Op. cit. p. 139ff.

> It appears that this resetting of the context has to rely on explicit indications [which are always non-conditional sentences in von Fintel's examples], whereas expansion occurs silently and smoothly.[114]

However, the above described conflict between (1) and (4) suggests that a conditional may shrink the horizon. Another line of reasoning, already mentioned in the section 6.3, leads to the same conclusion. The third premise, $A_3 \rightarrow D_4$, as we said, assumes $\neg C_3$. In terms of the metalinguistic theory, $\neg C_3$ is therefore one of the background propositions for the third conditional. However, that contradicts the first premise: $A_3 \rightarrow C_3$. If C_3 is a background proposition for the third conditional, than it has to be cotenable with A_3, i.e. $\neg(A_3 \rightarrow \neg\neg C_3)$, which is equivalent to $\neg(A_3 \rightarrow C_3)$, and that is the negation of the first premise. Therefore, asserting the third premise does involve a contradiction in the modal horizon, first because it implies the negation of the first premise, second because it assumes $\neg C_3$, which is a negation of the antecedent of the second premise.

At the end of the section 6.3 on CSO I concluded that the counterexample is not obviously good nor does it obviously fail, and that explanations for both have their good sides. Now I think that we can explain why this is so – because two interpretations of the conditionals involved are allowed: one that does not oppose CSO and the other that does. We have a conflict between (1) and (4). (1) forbids and (4) requires shrinking. However, it is not clear from the given context which one should prevail, (1) or (4). Von Fintel said that shrinking has to rely on *explicit* indications. In this case we have an *implicit* indication, i.e. we do not have an explicit contradiction in the horizon of the form A and ¬A, but a contradiction that has to be derived. Thus we can go two ways. We can avoid the contradiction by rejecting the third premise, as von Fintel and other supporters of CSO would do, but we can also avoid the contradiction by resetting the context (shrinking the horizon), which allows us to accept the third premise, as Gabbay would say we should do.

---

[114] Ibid. p. 140.

Let us turn now to the other counterexample to CSO[115]:

6.16 If I pressed the button, the light would be on. $(A \rightarrow C)$

6.17 If the light were on, I would press the button. $(C \rightarrow A)$

6.18 If I pressed the button, I would hear the anthem. $(A \rightarrow D)$

But not

6.19 If the light were on, I would hear the anthem. $(C \rightarrow D)$

Again the third premise carries the assumption that the light is off (¬C), contrary to the antecedent of the second conditional, and it contradicts the first premise, since for the third premise to be true, ¬C has to be cotenable with A. Thus again we have similar disagreement between von Fintel and Gabbay, but there is a new issue involved as well. As von Fintel does not address the problem of fragility, it is not clear whether he would allow the first two premises to be both true. There must be a time gap between the antecedent-event and the consequent-event, so either the two A's or the two C's from the first two conditionals are not the same. In any case, one of the two conditionals has to be backtracking, and hence false.

---

[115] The context was this. Pressing the button on my lamp turns the light on or off. When the light goes on, the lamp plays the anthem for a minute. This does not happen when the light goes off. It's around midnight, I want to sleep, so the light would bother me. The light is off.

In the section 6.3 we saw that the claim that one of A→C and C→A has to be backtracking, as well as the claim that either the premise or the conclusion of contraposition has to be backtracking, assumes fragility of the propositions involved. But this assumption seems to be bad, first because fragile propositions are something that we rarely encounter in ordinary speech, and second because the assumption trivializes the claims that CSO is valid and that contraposition is invalid. The assumption makes the premises of CSO always false, so CSO is trivially valid, and it makes the conclusion of contraposition trivially false, whenever the premise is true. Counterexamples to contraposition were a big discovery and a big surprise. They are not intended to say something trivial. The inventors of the axiom CSO and the first counterexamples to contraposition, Stalnaker and Lewis, do not treat the involved propositions as fragile in their arguments and examples. They treat the involved propositions, as it is natural to do in ordinary language, as non-fragile (Lewis is especially careful in that regard).

In the above counterexamples we do not interpret the antecedents and the consequents as fragile but as vague propositions, and we interpret the conditionals as non-backtracking. This is what we do in most cases of CSO and contraposition. Can we find a theoretical justification of such a practice? The first justification is that this practice is already a practice that cannot be changed, and *ipso facto* does not require a justification but only an explanation. Thanks to von Fintel, and using the terminology from this section, I think that we can offer an explanation. A→C and C→A and ¬C→¬A can all be non-backtracking because (4) (charity) influences (3) (fragility). First, A→C is asserted. The meaning of A and C and A→C and the truth value of A→C are determined by the previous context. If the previous context is not enough, or if we are at the beginning of a conversation, in accordance with the present state of affairs (2) and according to (4) we decide about fragility of A and C (3) so to make A→C true (assuming that this is possible). Then, C→A is asserted. As von Fintel taught us, C→A first changes the present context, and is to be evaluated in the new context. If this is consistent with the previous conversation, by (4) we determine fragility (3) of A and C in a way that makes C→A true (and therefore non-backtracking). Then we processed similarly when ¬C→¬A is asserted. It is therefore possible to interpret the three conditionals as true, or at least as non-backtracking.

To conclude, I think that both counterexamples to CSO can be interpreted in a way so that they work. In the case of the first counterexample, we had a conflict between (1) and (4). The conflict was over (2) – how to interpret the selection function or modal horizon? In the second counterexample we have a conflict of (1) and (4) over (3) – shall we accept throughout the conversation the interpretation of A and C exactly the same as in the first conditional, or shall we allow A and C to slightly change in order to avoid backtracking? In the first case both opposing intuitions – for and against CSO – were strong, because we didn't have a rule to decide between (1) and (4). An explicit contradiction would decide in favour of (4), but we only had an implicit contradiction. In the second case, the intuition against CSO prevails once we decide that the propositions are not fragile. But the degree of fragility was not fully determined by the given context, so both an interpretation in favour and an interpretation against CSO are possible. Of course, this means that CSO is not generally valid.

7.4 A modification of von Fintel's theory

There is something static in von Fintel's dynamic semantics, and this is the ordering of worlds. The ordering is fixed at the beginning of conversation, after the first conditional was asserted. The modal horizon expands to include the closest antecedent worlds and all the worlds more similar than those, or it remains a singleton containing only the actual world if the antecedent is true. In any case, some ordering is established and it doesn't change, no matter whether the modal horizon expands or shrinks. When the horizon expands, this doesn't mean that some worlds that were more remote became closer now. It means that the selection function selects more worlds. Some more remote worlds that were not taken into account before are taken into account now, but they stay equally remote as they were before.

Von Fintel does not say much about the similarity relation at the beginning of a discourse. He just borrows the initial truth conditions from 'standard theories' i.e. from Stalnaker and Lewis, without explaining what he means by similarity. Although similarity ordering is context-sensitive in the standard theories, von Fintel apparently does not think that our conversation can have any influence on those features of context that are relevant for the ordering. Stalnaker and Lewis, who tend to separate pragmatics from semantics, and who do not want similarity to be too sensitive to changes in context, may accept such an approach. But von Fintel is doing the opposite. He brings pragmatics into semantics. He makes truth values of conditionals dependent on previous conversation. For that reason I think that it is incumbent on him to say something about what determines the similarity relation. In particular, it seems to me that his semantics opens up the question: Why is it that the truth values of counterfactuals depend on the previous conversation, but similarity does not?

Anyway, since I reject the notion of similarity altogether, I will not deal with the questions that involve that notion. It is clear from the rest of this thesis that I would disagree with von Fintel on the question of the initial truth conditions. His dynamic semantics inherits the main problems that other theories based on absolute similarity have. Thus I would like to keep von Fintel's main contribution (the dynamic approach), and reject what he borrows from others. I would also like to include fragility, and to keep the pragmatic approach by distinguishing a pragmatic part and an underlying logic (as explained in 7.1). So I will continue my modification of Gabbay/ Warmbrōd/von Fintel in three steps. First I will try to make a dynamic version of Gabbay's theory. Then I will include the fragility of propositions, and finally discuss notions of validity and entailment that would explain examples used throughout this thesis. My ultimate goal is to involve the relevance logic as the underlying logic, but this is left to be done after finishing this thesis. The semantics I will propose here will be based on strict implication, although I believe that a kind of relevant implication is a better basis.

In the initial context a counterfactual $\Phi_1 \rightarrow \Psi_1$ is to be evaluated Gabbay's way:

7.25 $\Phi_1 \rightarrow \Psi_1$ is true at w iff $w \models \Phi_1 \rightarrow \Psi_1$ iff $\forall w' \in g(\Phi_1, \Psi_1, w)\ (w' \models \Phi_1 \supset w' \models \Psi_1)$

Thus '→' behaves as a strict implication on the domain of worlds selected by the function g. Φ_1 is brought within our modal horizon, and that must be taken into account by the truth conditions for any subsequent conditional. The function g should 'remember' what has previously been said. Let us add another parameter c_i as a fourth argument of the function g. c_i keeps track of the i-1 previous propositions that might have had influence on the modal horizon. Now the next conditional $\Phi_2 \rightarrow \Psi_2$ is asserted.

7.26 $\Phi_2 \rightarrow \Psi_2$ is true at w in the context c_2, i.e.

$w, c_2 \models \Phi_2 \rightarrow \Psi_2$ iff

$\forall w' \in g(\Phi_2, \Psi_2, w, c_2)\ (w' \models \Phi_2 \supset w' \models \Psi_2)$

What does the parameter c_2 do? It modifies the set determined by the ternary function $g(\Phi_2, \Psi_2, w)$ and turns it into the set $g(\Phi_1 \wedge \Phi_2, \Psi_2, w)$. Thus c_2 ensures that Φ_1 remains considered as a possibility. However, this is not good enough. The antecedents involved in our conversation may contradict each other, but this does not necessarily mean that our modal horizon contains a contradiction. 'If I win a lottery we will go to Spain; if not, we'll stay at home' is a sequence of conditionals that we want to be dynamically consistent. Our modal horizon considers both possibilities – winning and not winning – but not at the same worlds. Thus the role of c_2 should be this. If Φ_1 does not contradict Φ_2, then $g(\Phi_2, \Psi_2, w, c_2) = g(\Phi_1 \wedge \Phi_2, \Psi_2, w)$. Otherwise $g(\Phi_2, \Psi_2, w, c_2) = g(\Phi_2, \Psi_2, w)$.

Now $\Phi_3 \rightarrow \Psi_3$ is asserted.

7.27 $\Phi_3 \rightarrow \Psi_3$ is true at w in the context c_3, i.e.

$w, c_3 \models \Phi_3 \rightarrow \Psi_3$ iff

$\forall w' \in g(\Phi_3, \Psi_3, w, c_3)\ [w' \models \Phi_3 \supset w' \models \Psi_3]$

Similarly to the role of c_2, c_3 has the task of 'remembering' the previous two conditionals. $g(\Phi_3, \Psi_3, w, c_3) = g(\Phi_1 \wedge \Phi_2 \wedge \Phi_3, \Psi_3, w)$, provided that $\Phi_1 \wedge \Phi_2 \wedge \Phi_3$ is not a contradiction. Otherwise, the contradiction should be avoided. We have several possibilities.

Antecedents that contradict Φ_3 should not be included in the conjunction. For example, if Φ_1 does not contradict Φ_3 but Φ_2 does, then $g(\Phi_3, \Psi_3, w, c_3) = g(\Phi_1 \wedge \Phi_3, \Psi_3, w)$. If Φ_1 and Φ_2 contradict each other, but are both consistent with Φ_3, then $g(\Phi_3, \Psi_3, w, c_3) = g(\Phi_1 \wedge \Phi_3, \Psi_3, w) \cup g(\Phi_2 \wedge \Phi_3, \Psi_3, w)$.

7.28 **(Truth Conditions)** In general, $\Phi_n \rightarrow \Psi_n$ is true at the world w in the context c_n iff Ψ_n holds at all Φ_n-worlds from $g(\Phi_n, \Psi_n, w, c_n)$.

Let A be the set of all the antecedents Φ_i for $i \in [1, ..., n]$; let $A' = \{A_1, ..., A_m\}$ be a subset of the power set P(A); A' collects all consistent sets of antecedents from A that contain Φ_n and are not proper subsets of any consistent set from P(A); let $\wedge_j$ be a conjunction of all the propositions from A_j, for some $j \in [1, m]$ and $A_j \in A'$.

Then $g(\Phi_n, \Psi_n, w, c_n) = g(\wedge_1, \Psi_n, w) \cup ... \cup g(\wedge_m, \Psi_n, w)$.

Semantics based on 7.28 is similar to von Fintel's in some respects: transitivity is valid if the premises are in the order $\Phi \rightarrow \Psi$, $\Psi \rightarrow X$; it's invalid otherwise; CSO is valid; contraposition is invalid. It seems that the two semantics agree about the validity of the rules that involve only counterfactuals (i.e. where the premises and the conclusion are counterfactuals). This indicates that the two semantics capture the phenomenon of the dynamic meaning of counterfactuals in a similar way. This makes 7.28 similar to the 'standard theories' as well, since von Fintel's semantics validates most of the 'standard' rules. The differences seem not to be in the treatment of *sequences of conditionals*, but are caused more by the fact that the two semantics assign different truth values to particular counterfactuals, since one is based on a Gabbay-style selection function and the other on similarity. Thus in 7.28 we cannot have two unrelated contingent truths counterfactually implying each other, as it is in the 'standard theories'. 7.28 does not have problems with intuitively false conditionals of the form (small change) → (not big change), as in Anderson's example that I cited in Chapter 3.[116]

The fact that the two semantics assign different truth values to particular counterfactuals leads to disagreement about the validity of some formulae and rules. CEM is invalid in 7.28, which makes it different from Stalnaker's theory. CS is invalid in 7.28, which makes it different from both Stalnaker's and Lewis's theories.

CS $(\phi \wedge \psi) \supset (\phi \rightarrow \psi)$

In von Fintel's theory CS might fail if $\neg\psi$ is already in the modal horizon; otherwise it holds.

Another important similarity between 7.28 and von Fintel's theory is that both are made for cases where the modal horizon does not shrink. Besides that, both take into account the changes of the modal horizon caused by counterfactuals and no other propositions. We are therefore very far away from a general result. I will try to draw more general conclusions in the next section.

7.5 A pragmatic theory of counterfactuals

For a more general theory we need to take into account other aspects of context that have influence on the truth values of conditionals. Let us start with fragility.

---

[116] 3.31 If one were to scare a pregnant guinea pig, then all her babies would be born without tails.

7.29 $\Phi \rightarrow \Psi$ is true at the world w in the context c, i.e.

$$w, c \models \Phi \rightarrow \Psi \text{ iff}$$

$$w, c \models (\forall \phi \in \Phi^c)(\exists \psi \in \Psi^c)(\phi \rightarrow \psi) \text{ iff}$$

$$(\forall \phi \in \Phi^c)(\exists \psi \in \Psi^c)(\forall w' \in g(\Phi, \Psi, w, c))(w' \models \phi \supset w' \models \psi)$$

c is the context after the conditionals was asserted (not before). The meaning of Φ and Ψ may change from what it was before the conditional $\Phi \rightarrow \Psi$ was asserted, and their new meaning is Φ^c and Ψ^c. As explained in section 7.3, propositions are defined as sets of fragile propositions. Capital Greek letters stand for propositions and small letters for fragile propositions.

The function $g(\Phi, \Psi, w, c)$, as defined in 7.28, is not good enough for general purposes, because the modal horizon may shrink. We saw in section 7.3 that counterfactuals may cause the modal horizon to shrink. Besides that, non-conditional propositions have been ignored so far (in both von Fintel's semantics and my version of it), but they can have an influence on the modal horizon as well. They seem to work more like assumptions than like antecedents. This means that if they are accepted as claims, they should hold throughout the modal horizon (unlike antecedents which hold throughout the horizon only if they are tautologies). For example, if we have a non-shrinking sequence $\{\ldots, X, \Phi \rightarrow \Psi, \ldots\}$ where X is a non-conditional proposition, then X should hold throughout the set $g(\Phi, \Psi, w, c)$. X keeps on holding until we decide to rule it out, i.e. until the horizon shrinks. Of course, non-conditional propositions can shrink the horizon. For example, X in the above sequence would shrink the horizon if $\neg X$ were considered as a possibility before or if it were the antecedent of some previous conditional.

The function g should be defined to take into account the interaction of the four features of context from section 7.3: (1) the previous conversation, including all kinds of propositions, (2) the relevant states of affairs in the world, (3) fragility, (4) charity. The definition in 7.28 ignores (3) and (4) and some aspects of (1). As we said in section 7.3, the four factors sometimes do not determine meaning and do not eliminate the vagueness to a degree necessary to avoid misunderstandings (as in the thermometer example); sometimes, they might have contrary influence (as in the counterexamples to CSO). I believe that the four features are enough to eliminate all misunderstandings and imprecision in our conversation, but only if we prolong our conversation long enough with that goal. For example, had we kept on talking in order to clarify the content of the modal horizon and the fragility of propositions involved in the counterexamples to CSO, we could have decided which of the two proposed interpretations to accept. In general, we tolerate vagueness and imprecision until we encounter a problem in our communication. If we disagree about the truth value of a counterfactual, our disagreement may or may not be caused by a problem in communication. We can disagree about what the political situation in Rome would have been had Caesar not crossed the Rubicon, but this is likely to be a disagreement about the historical facts, our views on human nature, and the like. This is not a misunderstanding or a problem in communication, and therefore not an indication that the meaning is not determined enough by (1), (2), (3), and (4). However, if you claim that Cesar would have used nuclear weapons had he been in command of the USA army in North Korea, and I claim that he would have used catapults, this is not a disagreement about facts but primarily a problem in communication. Here we attach different meanings to the very same words:

the antecedent 'Caesar is in command in North Korea' has one meaning in your counterfactual and another meaning in mine. Different sets of fragile propositions would stand for your antecedent and for mine (3), and consequently the relevant background propositions (2) would be different.

All this presents difficulties for the formal treatment of the function g. Not only are (1), (2), (3), and (4) complicated by themselves, but they might leave room for different interpretations of meaning, which leads to different truth value assignments (as in the thermometer example), and they can be in a collision (as in the counterexamples to CSO) which again leads to disagreement about truth values. *We do not have rules in ordinary language that would resolve vagueness and misunderstandings of this kind.* The only way to remove them is to keep talking, until we get convinced that everybody taking part in the conversation is using the words in the same way. Charity (4) will lead us to correct our interpretations in order to avoid misunderstandings, or it will lead us to convince others to change their interpretations. (As long as we use vague propositions, we can never *know* that others are using the same interpretation that we do; but we assume that this is the case, and we hold that assumption until we encounter some problems in communication again. Vague propositions can be reduced to exact or fragile propositions, if we prolong our conversation long enough with the goal to successfully remove all the vagueness. However, as I said, we do not do that often.)

Thus the function g has to expand the modal horizon and to shrink it; to determine and to change the fragility of propositions; to determine the set of relevant facts about the world and to revise it later; to give us more than one interpretation and more than one truth value for a proposition when (1), (2), (3), and (4) allow misunderstandings or when they are in a collision. I do not know whether it is possible to define g formally. I do not say that it is impossible either. Actually, I think that something useful would come from an attempt to do that, no matter whether the final goal would be achieved or not. But I will not do it in this thesis. I will try something else.

We can keep on using the function g(Φ, Ψ, w, c) without specifying formally the role of the parameter c. Instead, we can use an informal description of the behavior of g with respect to c, which is basically given in section 7.3 and in the previous few paragraphs. This is enough to tell us what c might do – it can shrink/expand the horizon, change fragility, revise the set of relevant facts. Then on the meta-level we can consider some restrictions of the form: if c does this, then the notion of entailment would be...

Let us see first what happens if we do not put on any restrictions. Then, among other things, the modal horizon can shrink. If this is so, then, potentially, the horizon may shrink to the initial context. In that case any new conditional in our conversation can be evaluated independently, as if it was not a part of the conversation. This leads to a collapse to Gabbay's original 'entailment-free' logic – almost no rules of inference involving counterfactuals hold any more. This can also be seen as a justification of Gabbay's logic. The fact that his logic is so weak is not just his mistake. There is a deeper reason for a logic of conditionals to be weak, and this is the context-dependence of conditionals. Counterfactuals are extremely sensitive to context, and they also have influence on context. Two (or more) counterfactuals can often have incompatible influences on context, i.e. they do not 'go together'. This shift of context is the reason why all these counterexamples to most of the rules of inference involving counterfactuals are possible.

On the other hand, we do use many rules in ordinary speech, even though counterexamples are possible. In using these rules we believe that we are rational. Moreover, we believe that it would be irrational to reject our inferences. This means that we assume some restriction on the potential that counterfactuals have to influence the context. There are at least two restrictions which lead to two notions of entailment and which I think are useful for the purpose of explaining our ordinary language practice.

We can consider the contexts where the modal horizon does not shrink. This is the first restriction I have in mind, and it gives us the von Fintel-style logic that I defined in section 7.4 and modified in this section (the truth conditions are given in 7.29 and the function g is defined the same as in 7.28). It is possible that this logic is equivalent to Stalnaker's logic minus his last two axioms listed in Chapter 3 (page 25) – CV and CEM (which is the same as Lewis's logic minus CV and CS). This is a very weak logic, but considerably stronger than Gabbay's.

In section 7.2 we saw that besides the notion of dynamic validity von Fintel proposed another notion of entailment according to which an argument from $\phi_1, \ldots, \phi_n$ to ψ is valid iff

7.17 for all contexts *c* such that $c = c|\phi 1| \ldots |\phi n|$,

$$[\phi_1]^{c} \cap [\phi_2]^{c|\phi 1|} \cap \ldots \cap [\phi_n]^{c|\phi 1| \ldots |\phi n\text{-}1|} \subseteq [\psi]^{c|\phi 1| \ldots |\phi n|}$$

This notion is supposed to capture the notion of validity of 'logical arguments'. Von Fintel did not explain what exactly he meant by the notion of a logical argument, but he said that in logical argumentation we are committed to a stable context ('In classical logic, it is considered an imperative that in the assessment of arguments the context remain stable'[117]). After everything I said in this Chapter, it is probably obvious that I do not think that counterfactuals are a convenient tool for defining logical argumentation. A proper tool would rather be what I called in section 7.1 'underlying logic', i.e. a context-free logic. A 'stable context' is not something that goes with counterfactuals, since they are not only context dependent, but they influence context as well. Thus I do not find the notion of entailment in 7.17 helpful simply because the restriction $c = c|\phi 1| \ldots |\phi n|$ is satisfied extremely rarely.

We cannot hope to keep the context fixed for counterfactuals, but we can focus on those cases where counterfactuals do not influence the context in such a way to change the truth values of other counterfactuals. This is the second restriction that I propose. Besides the von Fintel-style notion of entailment described above, we need another one to explain the cases where we use much more rules of inference, including some monotonic rules like transitivity and contraposition. The second notion of entailment is Warmbrod-style. First, we can ignore the phenomenon of the dynamic nature of counterfactuals that von Fintel talked about, and evaluate conditionals as if they occurred in an initial context (maybe that it also what von Fintel had in mind in his 7.17). That way the order of the premises is not important any more. Then we can put a restriction that would not allow the conditionals to 'mess' with each other, i.e. we can consider only conditionals that 'go together', in the sense defined in section 7.1:

7.1 counterfactuals *go together* only if each of the background propositions for any of the counterfactuals is cotenable with each of their antecedents.

---

[117] von Fintel 2001 p. 141.

Now we can express 7.1 using the terminology that we developed in the meantime and give an improved version. Let B_i be the set of all the relevant background propositions for the counterfactual $\Phi_i \rightarrow \Psi_i$. The members of B_i are then true at all the worlds from $g(\Phi_i, \Psi_i, w, c_i)$, and there is no other world where all propositions from B_i are true. Let us also introduce a notion of cotenability of a set of propositions with a proposition. A set B_i is cotenable with the proposition Φ iff it is not the case that B_i would not have been a set of true propositions had Φ been true.

7.30 Counterfactuals $\Phi_1 \rightarrow \Psi_1, \ldots, \Phi_n \rightarrow \Psi_n$ *go together* iff the set B is cotenable with each of the antecedents $\Phi_1, \ldots, \Phi_n$, where $B = B_1 \cup B_2 \cup \ldots \cup B_n$.

However, 7.30 is not good enough, because it uses the notion of cotenability that does not make sense when the truth conditions are defined as in 7.29. The cotenability from 7.30 can be formulated within an absolute or an antecedent-relative theory, but not in a theory based on a Gabbay-style selection function, as it is the case in 7.29. The distinctive feature of Gabbay's logic – that the function g selects different worlds for counterfactuals that have the same antecedents but different consequents – has influence on the notion of cotenability as well, since cotenability is defined in terms of counterfactuals. In section 6.3 on CSO we saw that there might be a proposition B cotenable with A in the context of evaluating A→C, while in the context of evaluating A→D, B is not cotenable with A. We can find a true counterfactual A→C where both A and C are false. Then we can find a proposition D that is a logical consequence of A and ¬C and possibly some more true propositions such that A→D is true as well. (On that basis I made all those counterexamples to CSO.) For example (see page 80 above)

4.1 Had I struck m, it would have lit. (A→C)

6.2 Had I struck m, I would have lost my job $(A \rightarrow D)$

$\neg C$ (that the match doesn't light) is a background proposition for 6.2 and hence cotenable with A. But $\neg C$ obviously cannot be cotenable with A in the context of 4.1. Cotenability is therefore relative to the consequent. I will shortly say that 'X is Ψ-cotenable with Φ' meaning that a proposition X is cotenable with the proposition (antecedent) Φ relative to the proposition (consequent) Ψ. Thus $\neg C$ in the above example is D-cotenable with A, but it is not C-cotenable with A. Finally,

7.31 Counterfactuals $\Phi_1 \rightarrow \Psi_1, \ldots, \Phi_n \rightarrow \Psi_n$ *go together* iff the set B is Ψ_i-cotenable with Φ_i for each $i \in [1, n]$, where $B = B_1 \cup B_2 \cup \ldots \cup B_n$, and where B_i for each $i \in [1, n]$ is a set of all the propositions that are true at all the worlds from $g(\Phi_i, \Psi_i, w, c_i)$.

This completes my modification of Warmbrōd's theory.

Thus I propose truth conditions for counterfactuals as defined in 7.29, and I propose two notions of entailment, one dynamic with the function g defined as in 7.28, the other static based on 7.31. All these definitions assume that the antecedents are possible. To get a more general result, I stipulate that counterfactuals with impossible antecedents are true, and that the function g selects an empty set for such conditionals. This is the result that I do not like, and this one of the main reasons why I think that conditional logic should be based on relevance rather than modal logic. But I have to leave that for some other opportunity.

My point is this. The set of all true counterfactuals is inconsistent. This is the result we get if we define the notion of consistency in any logic stronger than Gabbay's

G. Counterfactuals cannot all go together, because they are context sensitive and they influence context, in the way explained in this chapter. This is why we cannot have a logic for counterfactuals, in the sense of 'logic' that we are used to. We need a pragmatic theory. To explain why rational people use some rules of inference, we need to discover some restrictions that tell us which counterfactuals go together. I propose two such restrictions. One is that counterfactuals can go together as long as they do not cause the modal horizon to shrink. The other is 7.31. My modifications of von Fintel's and Warmbrōd's logics give us the notions of entailment that correspond to these two restrictions.

Bibliography

Adams, Ernest W.

1975 **The Logic of Conditionals: An Application of Probability to Deductive Logic** Reidel Publishing Company, Dordrecht

1998 **A Primer of Probability Logic** CSLI Publications, Stanford

Anderson, Alan Ross and Belnap, Nuel D.

1975 **Entailment: The Logic of Relevance and Necessity** Princeton University Press.

Anderson, Alan Ross, Belnap, Nuel D. and Dunn, J. Michael

1992 **Entailment: The Logic of Relevance and Necessity** volume II, Princeton University Press.

Åqvist, Lennart

1973 "Modal logic with subjunctive conditionals and dispositional predicates" *Journal of Philosophical Logic* 2, pp. 1-76

Bennett, Jonathan

1974 "Counterfactuals and possible worlds" *Canadian Journal of Philosophy* 4, pp. 381-402

2003 **A philosophical guide to conditionals** Oxford University Press

Blue, N. A.

1981 A metalinguistic interpretation of counterfactual conditionals" *Journal of Philosophical Logic* 10, pp. 179-200.

Collins, J., Hall, N., and Paul, L. A. (eds.)

2004 **Causation and Counterfactuals,** MIT Press

Crocco, G., Farinas del Cerro, L., Herzig, A. (eds.)

1995 **Conditionals: from philosophy to computer science** Oxford University Press

Cross, Charles B.

1985 "Jonathan Bennett on 'even if' " *Linguistics and Philosophy* 8. pp. 353-357

Dunn, Michael

1986 "Relevance Logic and Entailment" in D. Gabbay and F Guenthner (eds.) Vol III. pp. 117-229

Edgington, Dorothy

1995 "On Conditionals" *Mind* 104. pp. 235-329.

Fetzer, J. H. and Nute, D.

1979 "Syntax, semantics and ontology: a probabilistic causal calculus" *Synthese* 40, pp. 453-495

1980 "A probabilistic causal calculus: conflicting conceptions" *Synthese* 44, pp. 241-246

von Fintel, Kai

2001 "Counterfactuals in a dynamic context" in M. Kenstowicz (ed.) pp.123-152

Gabbay, Dov M.

1972 "A general theory of the conditional in terms of a ternary operator" *American Philosophical Quarterly*, 3

1976 **Investigations in modal and tense logic with applications to problems in philosophy and linguistics** Reidel, Dordrecht

Gabbay, D. M. and Guenthner, F. (eds.)

1986 **Handbook of Philosophical Logic** 1st edition, Kluwer Academic Publishers

2002 **Handbook of Philosophical Logic** 2nd edition, Kluwer Academic Publishers

Goble, Louis (ed.)

2001 **Blackwell's guide to philosophical logic** Blackwell, Oxford

Goodman, Nelson

1947 "The Problem of Counterfactual Conditionals" *Journal of Philosophy*; 44: pp. 113-128.

1957 "Parry on counterfactuals", *Journal of Philosophy* 54, pp. 442-5.

1984 **Fact, Fiction, and Forecast** Cambridge, Mass. Originally published in 1954

1991 The same as Goodman 1947, reprinted in Jackson (ed.) 1991.

Grice, Paul

1967 "Logic and conversation" in Grice 1989

1989 **Studies in the way of words** Cambridge, Mass.: Harvard University Press

Hansson, Sven Ove

1995 "The emperor's new clothes: some recurring problems in the formal analysis of counterfactuals" in Crocco, G., Farinas del Cerro, L., Herzig, A. eds., 1995, pp. 13-31.

Harper, William L., Stalnaker, Robert, and Pearce, Glenn, (eds.)

1981 **Ifs: Conditionals, Belief, Decision, Chance, and Time** Dordrecht: Reidel

Heim, Irene and Kratzer, Angelika

1998 **Semantics in Generative Grammar** Blackwell

Henkin, Leon

1949 "The completeness of the first-order functional calculus" *Journal of Symbolic Logic*, 3, pp 159-166

Jackson, Frank (ed.)

1991 **Conditionals** Oxford University Press.

Kenstowicz, Michael (ed.)

2001 **Ken Hale: A life in language** MIT Press

Kratzer, Angelika

1979 "Conditional necessity and possibility" in R. Bauerle, U Egli, and A von Stechow (eds.) **Semantics from different points of view**, Sringer-Verlag, Berlin, 1979.

1981 "Partition and revision: the semantics of counterfactuals" *Journal of Philosophical Logic* 10, pp. 201-216.

Lange, Mark

1993 "When would natural laws have been broken?" *Analysis* 53 pp. 292-9

Lewis, David

1973 **Counterfactuals**, Cambridge, Harvard University Press.

1973a "Causation" *Journal of Philosophy* 70, pp. 556-67. Reprinted in Sosa (ed.) 1957 and in Lewis 1986a.

1986a **Philosophical Papers vol. II** Oxford University Press

1986b "Events" in Lewis 1986a pp. 241-69

1986c "Postscript to 'Causation'" in Lewis 1986a pp. 172-213

2000 "Causation as influence" *Journal of Philosophy* 97, pp. 182-97. An expanded version appears in Collins et al. (eds.) 2004.

Lowe, E. J.

1987 "Not a counterexample to modus ponens", *Analysis* 47

Mackie, John L.

1962 "Counterfactuals and causal laws" in R. J. Butller: **Analytical Philosophy** Blackwell, Oxford

Maslen, Cei

2004 "Causes, contrasts, and the nontransitivity of causation" in J. Collins et al. (eds.) 2004

Mares, Edwin D.

2004 **Relevant Logic: A Philosophical Interpretation** Cambridge University Press.

Mares, Edwin D. and Meyer, Robert K.

2001 "Relevant Logics" in L. Goble (ed.) 2001

Mårtensson, Johan

2000 **Subjunctive conditionals and time: a defence of a weak classical approach**

http://www.phil.gu.se/johan/johan.html

McKay, T. and van Inwagen, P

1977 "Counterfactuals with disjunctive antecedents" *Philosophical Studies* 31 pp. 353-6

McGee, Vann

1985 "A counterexample to modus ponens" *Journal of Philosophy* 9, pp. 462

Nute, Donald

1975a "Counterfactuals" *Notre Dame Journal of Formal Logic* 16. pp. 476-782

1975b "Counterfactuals and the similarity of worlds" *Journal of Philosophy* 72, pp. 773-778

1980 **Topics in conditional logic**, Reidel, Dordrecht

1981 "Causes, laws, and law statements" *Synthese* 48, pp. 347-370

Nute, Donald and Cross, Charles B.

2002 "Conditional Logic" in Gabbay, D. M. and Guenthner, F. (eds.) 2002, vol. 4, pp. 1-98

Parry, W. T.

1957 "Reexamination of the problem of counterfactual conditionals", *Journal of Philosophy* 54, pp. 85-94.

Perry, John

1986 "Thought without representation" *Proceedings of the Aristotelian Society* 60, pp. 137-51

Pollock, John L.

1976 **Subjunctive reasoning** D. Reidel Publishing Company, Dordrecht

1981 "A refined theory of counterfactuals" *Journal of Philosophical Logic* 10, pp. 239-266.

Quine, W.V.O.

1990 "Three Indeterminacies" in **Perspectives on Quine**, Robert B Barrett and Roger F. Gibson (Eds.), Blackwell, Oxford

Read, Stephen

1988 **Relevant Logic** Oxford (Basil Blackwell)

Rescher, Nicolas

1964 **Hypothetical reasoning** North-Holland: Amsterdam

1968 (ed.) **Studies in logical theory** Oxford, Blackwell

Sellars, W. S.

1957 "Counterfactuals" in Sosa (ed.) 1975 pp. 126-55

Sosa, Ernest (ed.)

1975 **Causation and Conditionals** Oxford University Press

Stalnaker, Robert

1968 "A theory of conditionals" in Rescher (ed.) 1968. Reprinted in Sosa 1975 and Harper et al. (eds.) 1981.

1978 "A defense of conditional excluded middle" in Harper et al. 1981.

Stalnaker, Robert and Thomason, Richmond

1970. "A semantic analysis of conditional logic" *Theoria* vol. 36. pp. 23-42

Turner, Raymond

1981 "Counterfactuals without possible worlds" *Journal of Philosophical Logic* 10. pp. 453-93.

van Fraassen, Bas C.

1966 "Singular terms, truth-value gaps, and free logic" *Journal of philosophy* 63, pp. 481-95

Warmbrōd, Ken

1981 "Counterfactuals and substitution of equivalent antecedents" *Journal of Philosophical Logic* 10, pp. 267-289

1982 "A defence of the limit assumption" *Philosophical Studies* 42, pp. 53-66

Veltman, Frank

1976 "Prejudices, presuppositions, and the theory of conditionals" in J. Groenendijk and M. Stokhof (eds.) **Amsterdam papers in formal grammar**, vol. 1 Centrale Interfaculteit, Universiteit van Amsterdam, pp. 248–281